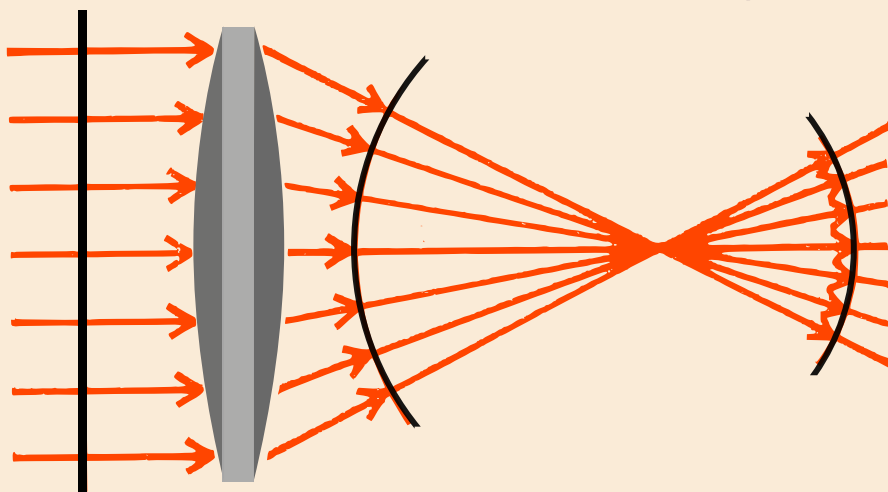


# ELEMENTARY TEXTBOOK ON PHYSICS

*Edited by  
G. I. Landsberg*

Volume 3



Mir Publishers Moscow

# ELEMENTARY TEXTBOOK ON PHYSICS

Edited by G.S. Landsberg

These three volumes form a course on elementary physics that has become very popular in the Soviet Union. Each section was written by an authority in the appropriate field, while the overall unity and editing was supervised by Academician G.S. Landsberg (1890-1957). This textbook has gone through ten Russian editions and a great deal of effort went into the last edition to introduce SI units and change the terminology and notation for the physical units.

A feature of this course is the relatively small number of formulas and mathematical manipulations. Instead, attention was focussed on explaining physical phenomena in such a way as to combine scientific rigour and a form understandable to school children. Another aspect of the text is the technological application of the physical laws.

These features make the text a world-class textbook.

For students preparing to enter universities and colleges to study physics, and for those in high schools specialising in physics.









ELEMENTARY  
TEXTBOOK ON  
**PHYSICS**

**Volume 3**

# ЭЛЕМЕНТАРНЫЙ УЧЕБНИК ФИЗИКИ

Под редакцией академика  
Г.С. ЛАНДСБЕРГА

в 3-х томах

ТОМ 3

КОЛЕБАНИЯ И ВОЛНЫ.  
ОПТИКА.  
АТОМНАЯ И ЯДЕРНАЯ ФИЗИКА

Издательство «Наука»  
Москва

# **ELEMENTARY TEXTBOOK ON PHYSICS**

---

**Edited by G.S. Landsberg**

---

In three volumes

Volume 3

**OSCILLATIONS AND WAVES  
OPTICS  
ATOMIC AND NUCLEAR PHYSICS**



Mir Publishers Moscow 

Translated from Russian  
by Natalia Wadhwa

First published 1989  
Revised from the 1986 Russian edition

*На английском языке*

*Printed in the Union of Soviet Socialist Republics*

ISBN 5-03-000264-2  
ISBN 5-03-000223-5

© Издательство «Наука». Главная редакция  
физико-математической литературы, 1986  
© English translation, Mir Publishers, 1989

# Contents

## Preface to the First Russian Edition 11

## Part One. Oscillations and Waves 13

### Chapter 1. Basic Concepts. Mechanical Vibrations 13

1.1. Periodic Motion. Period (13). 1.2. Oscillatory Systems. Free Oscillations (13). 1.3. Pendulum: Kinematics of Oscillations (15). 1.4. Vibrations of a Tuning Fork (16). 1.5. Harmonic Oscillations. Frequency (18). 1.6. Phase Shift (21). 1.7. Dynamics of Pendulum Oscillations (23). 1.8. Formula for the Period of a Simple Pendulum (25). 1.9. Elastic Vibrations (27). 1.10. Torsional Vibrations (29). 1.11. Effect of Friction. Damping (30). 1.12. Forced Vibrations (33). 1.13. Resonance (34). 1.14. Effect of Friction on Resonance Phenomena (36). 1.15. Examples of Resonance Phenomena (37). 1.16. Resonance Phenomena Induced by an Anharmonic Periodic Force (39). 1.17. The Relation Between the Form and Harmonic Composition of Periodic Oscillations (42).

### Chapter 2. Acoustic Vibrations 30

2.1. Acoustic Vibrations (30). 2.2. Subject of Acoustics (47). 2.3. Musical Tone. Loudness and Pitch (48). 2.4. Timbre (49). 2.5. Acoustic Resonance (51). 2.6. Recording and Reproduction of Sounds (53). 2.7. Analysis and Synthesis of Sound (54). 2.8. Noises (55).

### Chapter 3. Electric Oscillations 58

3.1. Electric Oscillations and Methods of Their Observation (58). 3.2. Oscillatory Circuit (61). 3.3 Mechanical Analogy. Thomson Formula (64). 3.4. Electric Resonance (67). 3.5. Undamped Oscillations. Self-Excited Oscillatory Systems (70). 3.6. Valve Oscillator (73). 3.7. Theory of Oscillations (75).

### Chapter 4. Wave Phenomena 79

4.1. Waves (79). 4.2. Wave Propagation Velocity (81). 4.3. Radiolocation, Hydroacoustic Detection and Sound Ranging (83). 4.4. Transverse Waves in a Cord (85). 4.5. Longitudinal Waves



in an Air Column (88). 4.6. Waves on the Surface of a Liquid (91). 4.7. Energy Transfer by Waves (93). 4.8. Reflection of Waves (96). 4.9. Diffraction (98). 4.10. Directional Emission (100).

## Chapter 5. Interference of Waves 103

5.1. Superposition of Waves (103). 5.2 Interference of Waves (104). 5.3. Conditions for Formation of Interference Maxima and Minima (106). 5.4. Interference of Acoustic Waves (108). 5.5. Standing Waves (109). 5.6. Vibrations of Elastic Bodies as Standing Waves (111). 5.7. Free Vibrations of a String (112). 5.8. Standing Waves in Plates and Other Extended Bodies (115). 5.9. Resonance in the Presence of Many Frequencies (117). 5.10. Conditions for a Perfect Sound Emission (119). 5.11. Binaural Phase Effect. Sound Direction Finding (121).

## Chapter 6. Electromagnetic Waves 123

6.1. Electromagnetic Waves (123). 6.2. Conditions for a Perfect Emission of Electromagnetic Waves (124). 6.3. Oscillators and Aerials (125). 6.4. Hertz' Experiments on Electromagnetic Waves. Lebedev's Experiments (129). 6.5. Electromagnetic Theory of Light. Scale of Electromagnetic Waves (132). 6.6 Experiments with Electromagnetic Waves (134). 6.7. Popov's Invention of Radio (141). 6.8. Modern Radio Communication (144). 6.9. Other Applications of Radio (147). 6.10. Propagation of Radio Waves (149). 6.11. Concluding Remarks (153).

## Part Two. Geometrical Optics 157

### Chapter 7. Light Phenomena: General 157

7.1. Effects of Light (157). 7.2. Interference of Light. Colours of Thin Films (159). 7.3. Brief Information from the History of Optics (160).

### Chapter 8. Photometry and Lighting Engineering 162

8.1. Radiant Energy. Luminous Flux (162). 8.2. Point Sources of Light (163). 8.3. Luminous Intensity and Illuminance (165). 8.4. Laws of Illumination (166). 8.5. Units of Photometric Quantities (168). 8.6. Brightness of Sources (169). 8.7. Problems of Lighting Engineering (171). 8.8. Appliances for Concentrating Luminous Flux (172). 8.9. Reflectors and Scatterers (173). 8.10. Brightness of Illuminated Surfaces (176). 8.11. Photometry and Measuring Instruments (177).

### Chapter 9. Basic Laws of Geometrical Optics 182

9.1. Rectilinearity of Wave Propagation (182). 9.2. Rectilinear Propagation of Light. Light Rays (183). 9.3. Laws of Reflection and Refraction of Light (188). 9.4. Reversibility of Light Rays (192). 9.5. Refractive Index (194). 9.6. Total Internal Reflection (197). 9.7. Refraction in a Plane-parallel Plate (200). 9.8. Refraction in a Prism (201).

## Chapter 10. Application of Reflection and Refraction of Light for Image Formation 204

10.1. Light Source and Its Image (204). 10.2. Refraction in a Lens. Focal Points (205). 10.3. Images of Points Located on the Principal Optical Axis of a Lens. Lens Equation (210). 10.4. Applications of the thin Lens Equation. Real and Virtual Images (212). 10.5. Image of a Point Source and of an Extended Object Formed by a Plane Mirror. Image of a Point Source Formed by a Spherical Mirror (216). 10.6. Focal Point and Focal Length of a Spherical Mirror (219). 10.7. Relation Between the Positions of a Source and Its Image on the Principal Optical Axis of a Spherical Mirror (220). 10.8. Methods of Preparation of Lenses and Mirrors (221). 10.9. Images of Extended Objects Formed by Spherical Mirrors and Lenses (222). 10.10. Magnification of Images Formed by Spherical Mirrors and Lenses (223). 10.11. Image Formation by Spherical Mirrors and Lenses (225). 10.12. Optical Power of Lenses (230).

## Chapter 11. Optical Systems and Errors 232

11.1. Optical System (232). 11.2. Principal Planes and Principal Points of a System (232). 11.3. Image Construction in a System (234). 11.4. Magnification of a System (234). 11.5. Drawbacks of Optical Systems (235). 11.6. Spherical Aberration (236). 11.7. Astigmatism (239). 11.8. Chromatic Aberration (240). 11.9. Confinement of Beam Cross Sections in Optical Systems (241). 11.10. Lens Aperture (242). 11.11. Brightness of Image (243).

## Chapter 12. Optical Instruments 246

12.1. Projection Optical Instruments (246). 12.2. Photographic Camera (248). 12.3. The Human Eye as an Optical System (250). 12.4. Optical Instruments Outfitting the Eye (252). 12.5. Magnifying Glasses (254). 12.6. Microscopes (256). 12.7. Resolving Power of Microscopes (258). 12.8. Telescopes (258). 12.9. Magnification of Telescopes (260). 12.10. Telescopes in Astronomy (261). 12.11. Image Brightness for Extended and Point Sources (265). 12.12. Lomonosov's Telescope (267). 12.13. Binocular Vision and Sensation of Depth. Stereoscopes (267).

## Part Three. Physical Optics 272

### Chapter 13. Interference of Light 272

13.1. Geometrical and Physical Optics (272). 13.2. Experimental Realization of Interference of Light (272). 13.3. Explanation of Thin Film Colours (276). 13.4. Newton's Rings (277). 13.5. Calculation of Wavelength of Light with the Help of Newton's Rings (279).

### Chapter 14. Diffraction of Light 282

14.1. Bundles of Rays and the Shape of Wave Surface (282). 14.2. Huygens' Principle (283). 14.3. Reflection and Refraction from the Viewpoint of Huygens' Principle (284). 14.4. Huygens'

Principle in Fresnel Interpretation (286). 14.5. Simple Diffraction Phenomena (287). 14.6. Explanation of Diffraction by Fresnel's Method (290). 14.7. Resolving Power of Optical Instruments (291). 14.8. Diffraction Grating (294). 14.9. Diffraction Grating as a Spectral Instrument (296). 14.10. Preparation of Diffraction Gratings (297). 14.11. Diffraction at an Oblique Incidence of Light on a Grating (297).

## **Chapter 15. Physical Principles of Optical Holography 299**

15.1. Photography and Holography (299). 15.2. Holographic Recording with a Plane Reference Wave (302). 15.3. Obtaining Optical Images by Reconstructing the Wave Front (305). 15.4. Holographing by Opposing Light Beam Method (308). 15.5. Application of Holography to Optical Interferometry (310).

## **Chapter 16. Polarization of Light. Transverse Nature of Light Waves 315**

16.1. Passage of Light Through Tourmaline (315). 16.2. Hypotheses Explaining Observed Phenomena. Polarized Light (316). 16.3. Mechanical Model of Polarization (317). 16.4. Polaroids (318). 16.5. Transverse Nature of Light Waves and Electromagnetic Theory of Light (318).

## **Chapter 17. Electromagnetic Spectrum 320**

17.1. Methods of Investigating Electromagnetic Waves of Different Wavelengths (320). 17.2. Infrared and Ultraviolet Radiation (321). 17.3. Discovery of X-rays (322). 17.4. Effects of X-rays (324). 17.5. X-ray Tube (325). 17.6. Origination and Nature of X-rays (326). 17.7. Scale of Electromagnetic Waves (327).

## **Chapter 18. Speed of Light 329**

18.1. First Attempts to Determine the Speed of Light (329). 18.2. Determination of the Speed of Light by Roemer (330). 18.3. Measurement of the Speed of Light by Rotating-Mirror Method (331).

## **Chapter 19. Dispersion of Light and Colours of Bodies 334**

19.1. State-of-the-Art in Chromatography Before Newton's Studies (334). 19.2. Main Discovery of Newton in Optics (334). 19.3. Interpretation of Newton's Observations (336). 19.4. Dispersion of Refractive Indices for Different Materials (337). 19.5. Complementary Colours (338). 19.6. Spectral Composition of Light Emitted by Various Sources (340). 19.7. Light and Colours of Bodies (341). 19.8. Absorption, Reflection and Transmission Coefficients (342). 19.9. Coloured Bodies Illuminated by White Light (342). 19.10 Coloured Bodies Illuminated by Coloured Light (343). 19.11. Masking and Unmasking (344). 19.12. Colour Saturation (345). 19.13. Colour of the Sky and Dawns (346).

## Chapter 20. Spectra and Spectral Regularities 349

20.1. Spectroscopic Instrumentation (349). 20.2. Types of Emission Spectra (350). 20.3. Origin of Different Types of Spectra (352). 20.4. Spectral Laws (353). 20.5. Spectral Analysis Using Emission Spectra (354). 20.6. Absorption Spectra of Liquids and Solids (357). 20.7. Absorption Spectra of Atoms. Fraunhofer Lines (357). 20.8. Investigation of Red-Hot Bodies. Blackbody (358). 20.9. Temperature Dependence of Red-Hot Bodies. Incandescent Lamps (360). 20.10. Optical Pyrometry (361).

## Chapter 21. Effects of Light 363

21.1. Action of Light on a Substance. Photoelectric Effect (363). 21.2. Laws of Photoelectric Effect (364). 21.3. Light Quanta (367). 21.4. Application of Photoelectric Phenomena (369). 21.5. Photoluminescence. Stokes' Shift (371). 21.6. Physical Meaning of Stokes' Shift (373). 21.7. Luminescent Analysis (373). 21.8. Photochemical Action of Light (374). 21.9. The Role of Wavelength in Photochemical Processes (375). 21.10. Photography (375). 21.11. Photochemical Theory of Vision (379). 21.12. Duration of Visual Sensation (381).

## Part Four. Atomic and Nuclear Physics 388

### Chapter 22. Atomic Structure 388

22.1. Atoms (388). 22.2. Avogadro's Constant. Size and Mass of Atoms (389). 22.3. Elementary Electric Charge (391). 22.4. Units of Charge, Mass and Energy in Atomic Physics (394). 22.5. Measurement of Mass of Charged Particles. Mass Spectrograph (395). 22.6. Electron Mass. Velocity Dependence of Electron Mass (398). 22.7. Einstein's Law (400). 22.8. Mass of Atoms. Isotopes (403). 22.9. Isotope Separation. Heavy Water (405). 22.10. Nuclear Model of Atom (407). 22.11. Energy Levels of Atoms (410). 22.12. Induced Emission of Light. Quantum Generators (415). 22.13. Hydrogen Atom. Peculiarities of Motion of an Electron in an Atom (419). 22.14. Many-Electron Atoms. Origin of Optical and X-ray Spectra of Atoms (423). 22.15. Mendeleev's Periodic System of Elements (424). 22.16. Quantum and Wave Properties of Photons (427). 22.17. Fundamentals of Quantum (Wave) Mechanics (433).

### Chapter 23. Radioactivity 441

23.1. Discovery of Radioactivity. Radioactive Elements (441). 23.2. Alpha-, Beta- and Gamma-Radiation. Wilson Cloud Chamber (443). 23.3. Methods of Detecting Charged Particles (448). 23.4. Properties of Radioactive Radiation (451). 23.5. Radioactive Decay and Radioactive Transformations (455). 23.6. Applications of Radioactivity (459). 23.7. Accelerators (459).

### Chapter 24. Atomic Nuclei and Nuclear Power 465

24.1. Nuclear Reactions (465). 24.2. Nuclear Reactions and Transformation of Elements (467). 24.3. Properties of Neutrons (468). 24.4. Nuclear Reactions Induced by Neutrons (470). 24.5.

Artificial Radioactivity (472). 24.6. Positron (474). 24.7. Application of Einstein's Law to Annihilation and Pair Formation (476). 24.8. The Structure of Atomic Nuclei (477). 24.9. Nuclear Energy. Energy Sources of Stars (480). 24.10. Uranium Fission. Chain Nuclear Reaction (483). 24.11. Application of Nondecaying Chain Fission Reaction. Atom and Hydrogen Bombs (488). 24.12. Nuclear Reactors and Their Application (490).

## Chapter 25. Elementary Particles 499

25.1. General Remarks (499). 25.2. Neutrino (500). 25.3. Nuclear Forces. Mesons (502). 25.4. Particles and Antiparticles (506). 25.5. Particles and Interactions (511). 25.6. Detectors of Elementary Particles (513). 25.7. Clock Paradox (518). 25.8. Cosmic Radiation (Cosmic Rays) (519).

## Chapter 26. New Achievements in Elementary-Particle Physics 523

26.1. Accelerators and Experimental Technology (523). 26.2. Hadrons and Quarks (528). 26.3. Quark Structure of Hadrons (537). 26.4. Quark Model and Formation and Decay of Hadrons (538). 26.5. Leptons. Intermediate Bosons. The Unity of All Interactions (542).

## Answers and Solutions 546

Part I. Oscillations and Waves 546

Part II. Geometrical Optics 548

Part III. Physical Optics 551

Part IV. Atomic and Nuclear Physics 552

Conclusion 557

Index 560

# Preface to the First Russian Edition

The third volume of the *Elementary Textbook on Physics* is devoted to an analysis of phenomena associated with oscillations and waves. The chapters dealing with mechanical vibrations and elastic waves are concluded by an investigation of acoustic phenomena. The chapters concerning electromagnetic oscillations and waves are naturally associated with an analysis of the most significant application of these problems, viz. radio physics and radio engineering starting from the invention of the radio by A.S. Popov. These studies lead to the ideas about the electromagnetic origin of light and to the fundamentals of physical (wave) and geometrical (ray) optics. This volume ends with a brief description of the phenomena associated with modern atomic theory. Like the previous volumes, this volume also contains a large number of problems whose solution plays an important role in grasping the text.

The chapters dealing with mechanical and electromagnetic oscillations, acoustics and radio physics were written by S.M. Rytov. The part devoted to geometrical optics was written by M.M. Sushchinskii (with the participation of I.A. Yakovlev), the section on physical optics was contributed by F.S. Landsberg-Baryshanskaya, while the section on atomic and nuclear physics was written by F.L. Shapiro.

As in the previous volumes, the authors endeavoured to present the physical essence of the phenomena in a comprehensible form. Once again, mathematical calculations occupy an insignificant place in the book. Only a limited number of questions are beyond the scope of the high-school curriculum. However, in order to explain the topics considered in the book clearly, we covered many aspects in greater detail than in high-school textbooks. It should be emphasized that the problems on atomic and nuclear physics cover a wider range of topics than is covered by the high-school course. While allowing for expansion, we thought it proper to discuss these topics in greater detail, at the same time confining ourselves to the most interesting subjects. However, a consistent description of all the problems considered in these sections is quite difficult, and hence this part of the

book does not bear a close resemblance to the structure of the remaining sections. It is our earnest hope that it will evoke a keen interest among the students towards these important problems of modern physics.

This is the concluding volume of the course on elementary high-school physics. The whole project involved some ten authors, and hence the editor's task was difficult and full of responsibility since the diversity of approach and manner of presentation of the different authors differed even though a common approach towards the general principles underlying the course was adopted. In spite of the editor's efforts, greater diversity of presentation remains than is perhaps desirable for a textbook. However, the participation of so many authors has the definite advantage that it allows for collaboration between experts in various fields, each knowing a particular field and the difficulties associated with presenting the relevant topics.

Therefore, it is my earnest hope that in spite of all its drawbacks, the course will be beneficial to all students and will improve their understanding of physics, thereby increasing their interest in science as a whole. On this note of optimism, I conclude my epic task of compiling this *Elementary Textbook on Physics*.

Moscow, March 1952

*G. Landsberg*

# Part One

## Oscillations and Waves

### Chapter 1

## Mechanical Vibrations

### Basic Concepts.

#### 1.1. Periodic Motion. Period

Among a large variety of mechanical motions occurring in nature, repetitive motions are often encountered. Any uniform rotation is a repetitive motion: during each revolution a point on a uniformly rotating body passes through the same positions as in the previous revolution, in the same sequence and with the same velocities remaining unchanged. If we watch the branches and stems of trees shaking from the wind, a ship rolling on waves, a swinging pendulum of a clock, the reciprocal motion of pistons and connecting rods of a steam engine or a Diesel, up and down strokes of the needle of a sewing machine, the alternating tides and ebbs, the movements of arms and legs during walking or running, or feel beats of pulse, the same feature characterizes all these motions, viz. the multiple repetition of the same cycle of motions.

In actual practice, the repetition is not always completely the same in any conditions. Sometimes, every new cycle very accurately reproduces the previous cycle (swinging of a pendulum, motion of parts of a machine operating at a constant speed), while in other cases the difference between consecutive cycles may be noticeable (tides and ebbs, shaking of tree branches, or motion of the parts of a starting or stopping machine). Deviations from a completely exact repetition are often so small that they can be neglected, and the motion can be assumed repetitive to a high degree of accuracy. In other words, the motion is assumed to be *periodic*.

*A repetitive motion is referred to as periodic when each its cycle exactly reproduces any other cycle.*

The duration of a cycle is called a *period*.

Obviously, the period of a uniform rotation is equal to the duration of one revolution.

#### 1.2. Oscillatory Systems. Free Oscillations

Bodies and devices which are capable of performing periodic motions by themselves play an extremely important role in nature and especially in engineering. Saying “by themselves” we mean that they are not forced to oscillate by external periodic forces. For this reason, such oscillations are known



as *free* oscillations on contrast to *forced* oscillations occurring under the action of periodically varying external forces.

If, for example, we periodically push a door and pull it back, it will be opened and closed, i.e. will be in a forced periodic motion. But the door will not move periodically by itself: if we push it and leave to itself, its motion will not be repetitive. The situation is different if we push or deflect a load suspended on a string from the vertical. It will start to *swing*, i.e. will by itself perform a periodic motion. Then we obtain free oscillations. Similarly, as a result of an initial push, oscillatory motions will be performed by water in a glass, by a load suspended on a spring, by a carriage of a motor car on its leaf springs, by a swing, by a metallic plate fixed at one end, by a string, by a compass needle, and so on (Fig. 1).

All such bodies or systems of bodies which can perform periodic motions or oscillations by themselves are called *oscillatory systems*. As was stated above, the oscillations performed by these systems without influence of external forces are termed *free* oscillations.

Oscillatory systems are encountered not only in various machines and mechanisms (like a clockwork mechanism). It will be shown later that most of sources of sound are oscillatory systems, and the propagation of sound in air is possible just because the air itself is a sort of oscillatory system. Moreover, besides mechanical oscillatory systems, there exist electromagnetic oscillatory systems in which electric oscillations forming the basis of entire radioengineering may occur. Finally, there are many hybrid, viz. electromechanical oscillatory systems which are employed in various fields of engineering.

We shall start from an analysis of a simple mechanical oscillatory system, viz. a pendulum, and up to Sec. 1.12, we shall deal with free oscillations.

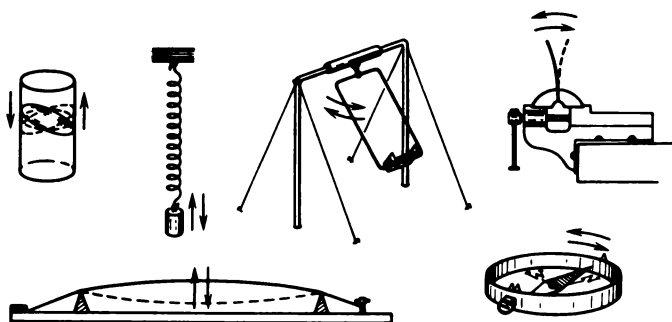


Fig. 1.  
Examples of free oscillations.

### 1.3. Pendulum Kinematics of Oscillations

Any body suspended so that its centre of gravity is lower than the point of suspension is a pendulum. A hammer suspended on a nail and a load on a string are examples of oscillatory systems resembling the pendulum of a wall clock (Fig. 2).

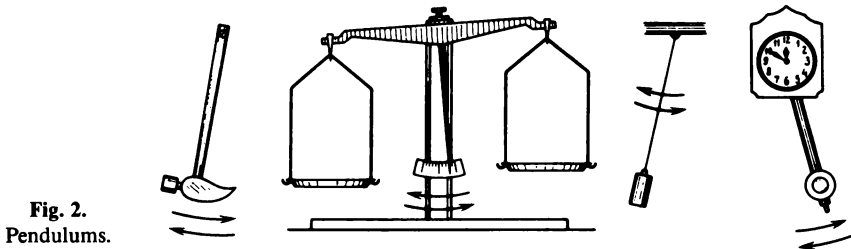


Fig. 2.  
Pendulums.

Any system capable of performing free oscillations is characterized by a stable equilibrium position. For a pendulum, this is the position in which the centre of gravity lies on the vertical below the point of suspension. If we draw the pendulum from this position or just push it, it starts to oscillate, being deflected to either side from the equilibrium position. The maximum deviation from the equilibrium position of the pendulum is called the *amplitude* of oscillations. The amplitude is determined by the initial deflection or by the push that sets the pendulum in motion. This property, viz. the dependence of the amplitude on the initial conditions, is typical not only of free oscillations of a pendulum, but in general of free oscillations of a large number of oscillatory systems.

Let us fix a hair (a piece of a thin wire or elastic nylon thread) to a pendulum and move a blackened glass plate under it as shown in Fig. 3. If we move the plate at a constant velocity in the direction perpendicular to the plane of oscillations, the hair will draw a wavy line on the plate

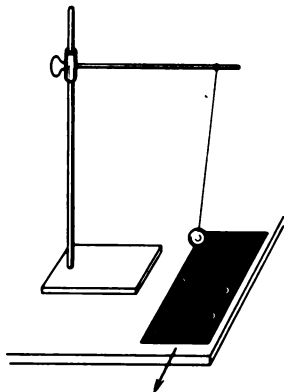


Fig. 3.  
Recording of pendulum oscillations on a blackened plate.

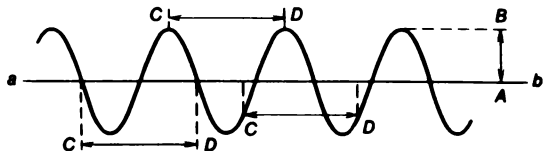


Fig. 4.  
Oscillogram of pendulum oscillations:  $AB$  is the amplitude and  $CD$  is the period.

(Fig. 4). In this experiment, we obtain the simplest *oscillograph*, viz. the instrument for recording oscillations. The curves recorded with the help of an oscillograph are known as *oscillograms*. Thus, Fig. 4 represents an oscillogram of oscillations of the pendulum. On this oscillogram, the amplitude of oscillations is represented by segment  $AB$  which gives the maximum deviation of the wavy curve from the straight line  $ab$  which would be drawn by the hair on the plate with a stationary pendulum (resting in the equilibrium position). The period is represented by segment  $CD$  equal to the distance covered by the plate during a period of the pendulum.

Since the blackened plate is moved uniformly, any its displacement is proportional to the time during which it has been completed. Therefore, it can be stated that time is laid off on a certain scale along the straight line  $ab$  (the scale depending on the velocity of the plate). On the other hand, the hair marks on the plate (in the direction perpendicular to  $ab$ ) the distances between the tip of the pendulum and its equilibrium position, i.e. the distance covered by the pendulum tip from this position.<sup>1</sup> Thus, the *oscillogram is just the graph of motion, viz. the curve expressing the dependence of the path length on time.*

It is well known that the slope of the curve on such a graph is equal to the velocity of motion (see Vol. 1, Sec. 1.19). The pendulum passes through the equilibrium position at the maximum velocity. Accordingly, the slope of the wavy curve in Fig. 4 is maximum at the points where it intersects the straight line  $ab$ . Conversely, the velocity of pendulum is equal to zero at the moments corresponding to maximum deviations. Accordingly, the tangent to the wavy line in Fig. 4 is parallel at the points of maximum separation from  $ab$  (i.e. the slope of the tangent is zero).

#### 1.4. Vibrations of a Tuning Fork

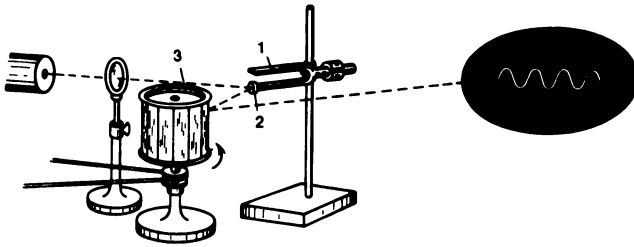
It was mentioned above that most of the sources of sound are oscillatory systems. We can easily verify that a sounding tuning fork vibrates, the form of its oscillations being the same as for a pendulum.

We can again take a blackened plate as an oscillograph fixing a recording hair to the prong of the tuning fork.

Since the amplitude and period of oscillations of the tuning fork are small, it is more convenient to use an oscillograph with a light spot and mirror scanning, which was described earlier (see Vol. 2, Sec. 17.2). Figure 5 illustrates the set-up for this experiment.

---

<sup>1</sup> To be more precise this is not the path length but rather the deviation of the pendulum from the equilibrium position. The displacement curve coincides with the path graph on almost straight segments between two adjacent extreme positions, and hence can be used for estimating the velocity of the pendulum. — *Eds.*



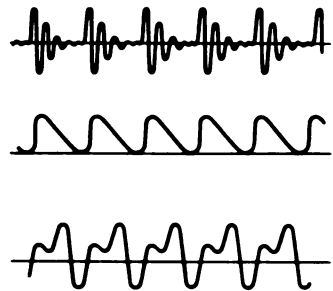
**Fig. 5.**  
Light-beam oscillograph with mirror scanning.

A small light mirror 2 is glued to a prong of a tuning fork 1. A light beam reflected at this mirror and at a mirror drum 3 forms a light spot (indicator) on the wall. If we strike the tuning fork we see that the spot is stretched into a vertical bright band. This is due to the fact that mirror 2 vibrates together with the prong of the tuning fork.

If we now set the drum in rotation, the light spot will receive a horizontal displacement, and the band will be scanned into a wavy line familiar to us from the previous experiment.

The amplitude and the period do not give a complete idea about the nature of a periodic motion. Indeed, a large variety of periodic motions may have the same period and amplitude but completely different forms of oscillations (forms of oscillograms). A few examples of oscillograms of such motions, representing the oscillations of certain mechanical and electric oscillatory systems, are shown in Fig. 6.

However, among a large variety of oscillations, the oscillations of a pendulum or of a tuning fork are of especial importance. The form of these oscillations is typical of a very large number of oscillatory systems. In particular, we can obtain the same oscillogram as for a pendulum by attaching a recording hair to a vibrating metal plate or to a load vibrating on a spring. The oscillogram of an alternating current represents the same form of oscillations (see Vol. 2, Sec. 17.3).



**Fig. 6.**  
Examples of oscillations with the same period, but different forms.

Therefore, it is necessary to consider the oscillations of this type in greater detail. In the next section, it will be shown that the oscillations of the form similar to that of a pendulum are connected in a simple way with the uniform motion in a circle. This will provide a method of graphic construction of the oscillogram for a pendulum.

### 1.5. Harmonic Oscillations. Frequency

Let us attach a ball fixed on a rod to a uniformly rotating disc and illuminate it from a side (Fig. 7). As the disc rotates, the shadow of the ball will oscillate on the wall. We can easily construct the graph of these oscillations. In Fig. 8, sixteen consecutive positions of the ball separated by  $1/16$  of a complete revolution are marked and numbered. The positions of the shadow on wall  $AB$  are numbered with the same figures from 1 to 16. These points are obtained by dropping perpendiculars from the points on the circle on the straight line  $AB$ . Just in this way the shadow is projected onto the wall when the ball is illuminated by a beam of parallel rays.

In order to scan the oscillations of the projection of the ball in the same way as it is done by a mirror drum, we construct a number of equidistant straight lines parallel to  $AB$ . The consecutive positions 1, 2, 3, . . . , 16 of the projection (shadow) will be plotted not on the same straight line but on the consecutive equidistant lines as shown on the right-hand side of Fig. 8. Connecting the thus obtained points by a continuous curve, we obtain a wavy line corresponding to consecutive positions of the shadow of the ball, i.e. the graph of motion. Thus, we obtain an "oscillogram" of oscillations of the projection of the ball.

The oscillation performed by the projection of a point moving uniform-

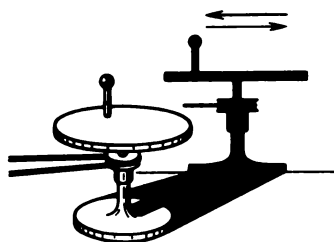


Fig. 7.  
Shadow projection of a ball moving in a circle.

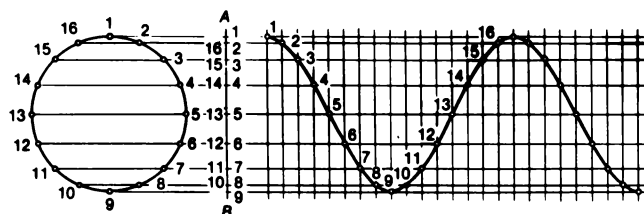


Fig. 8.  
Construction of the scan of a harmonic oscillation.

ly in a circle on a straight line is known as the *harmonic* (or simple) *oscillation*.

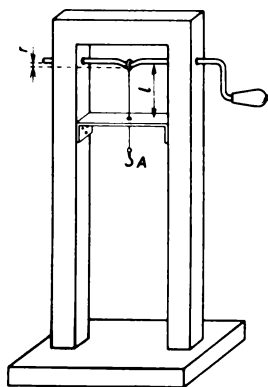
Harmonic oscillations form a special, particular type of periodic oscillations. This special type of oscillation is very important since it is often encountered in various oscillatory systems. The oscillations of a load on a spring, of a tuning fork, of a pendulum, or of a metal plate fixed at one end are examples of harmonic oscillation. It should be noted that for a large amplitude the oscillations in the above systems have a somewhat more complicated form, and the larger the amplitude, the less harmonic the oscillation.

An oscillation very close to a harmonic oscillation can be obtained with the help of a device shown in Fig. 9. When the handle is uniformly rotated, point  $A$  on a stretched string periodically moves up and down. If length  $l$  of the segment of the string to the orifice is large in comparison with sag  $r$  of the shaft, the motion of point  $A$  is very close to a harmonic oscillation. We shall use this simple device later.

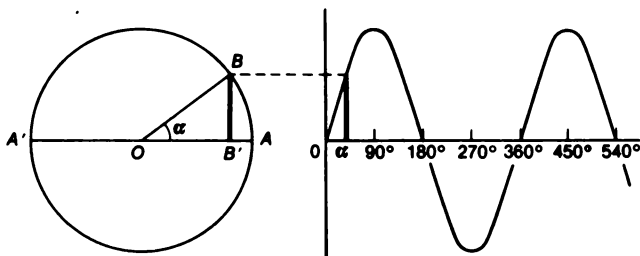
It should be noted that in the definition of harmonic oscillation a parallel projection has been mentioned. In other words, the positions of the point moving in a circle are plotted on the straight line  $AB$  (Fig. 8) by means of perpendiculars to  $AB$ , which are parallel to one another.

If we plot on the abscissa axis the central angle  $\alpha$  (Fig. 10) and on the

**Fig. 9.**  
Mechanism for obtaining harmonic motion.



**Fig. 10.**  
Construction of a sinusoid.



ordinate axis perpendicular  $BB'$  dropped from the tip of the rotating radius  $OB$  on the stationary diameter  $AA'$  (angle  $\alpha$  is measured from the fixed radius  $OA$ ), we obtain a curve known as a *sinusoid*. For each abscissa  $\alpha$ , the ordinate  $BB'$  of this curve is proportional to the sine of the angle  $\alpha$  since

$$\sin \alpha = \frac{BB'}{OB}.$$

Comparing this construction with the scanning of the harmonic oscillation described above, we can easily see that they are completely identical. Thus, the “wavy curve” representing a harmonic oscillation is a sinusoid. Therefore, a harmonic, or simple oscillation is often referred to as a *sinusoidal* oscillation.

The number of cycles of a harmonic oscillation completed during a second is called the *frequency* of this oscillation. If the period of a pendulum is 1 s (a second pendulum), one cycle is completed in a second, and the oscillatory frequency is equal to unity. The unit of frequency is termed *hertz* (Hz) after the German physicist H. Hertz (1857-1894) who obtained electrical oscillations (see below). As usual, the prefixes “kilo” and “mega” denote thousand and million times larger units:

$$1 \text{ kilohertz} = 1000 \text{ hertz}$$

$$1 \text{ megahertz} = 1\,000\,000 \text{ hertz}.$$

If a period is equal to 5 s, the frequency is 1/5 Hz. In general, denoting the duration of a period in seconds by  $T$  and the frequency in hertz by  $\nu$ , we obtain

$$\nu = \frac{1}{T}.$$

Thus, the period  $T$  of a harmonic oscillation determines the frequency  $\nu = 1/T$ . It should be remembered, however, that this relation between the period and frequency is typical of only harmonic (sinusoidal) oscillations. For a periodic oscillation of a different (nonharmonic) form, there is no definite frequency although it is characterized by a period  $T$ . This will be clarified later (see Sec. 1.17). Therefore, when we speak of an oscillation occurring at a certain frequency, we always mean a harmonic oscillation rather than a periodic motion of an arbitrary form.

In science and technology, we sometimes come across mechanical oscillations with frequencies varying over a wide range. For example, the pendulum for the demonstration of Foucault's experiment,<sup>2</sup> suspended at the

---

<sup>2</sup> This experiment makes it possible to detect the diurnal rotation of the Earth from the rotation of the plane of oscillations of the pendulum.

doom of the St. Isaac cathedral in Leningrad, has a period of about 20 s, i.e. its frequency is  $\nu = 0.05$  Hz. On the other hand, the frequency of vibrations of a railway carriage on its springs is about 1 Hz, while tuning forks may vibrate at frequencies of a few tens or hundreds of hertz. Physicists know how to obtain the so-called ultrasonic vibrations (which will be considered later) at frequencies reaching a few tens of megahertz. The atomic oscillations in molecules occur at a frequency of  $10^6$  Hz. Thus, the range of frequencies of mechanical vibrations is indeed very wide.

In these examples, we speak about frequency, thus indicating that these oscillations are *harmonic*.

### 1.6. Phase Shift

Can two harmonic oscillations having the same amplitudes and frequencies differ from each other? Let us take two *identical* pendulums and deflect them from the vertical through the same angle. If we now release them, we shall obtain two harmonic oscillations with the same amplitudes and frequencies. It appears that there can be no difference between them.

If, however, we release these pendulums not at once, we shall immediately notice the difference: the oscillations will be shifted in time.

Let us first release one pendulum and then the other at the moment when the first pendulum passes through the equilibrium position (i.e. in a quarter of a period). The pendulums will oscillate all the time with the same time shift by a quarter of a period (Fig. 11).

We could release the second pendulum half a period after the first pendulum has been released. The oscillations would be shifted by half a period: they pass through the equilibrium position simultaneously, but always move in opposite directions. When one of them is at the extreme right position, the other will be in the extreme left position, and vice versa (Fig. 12).

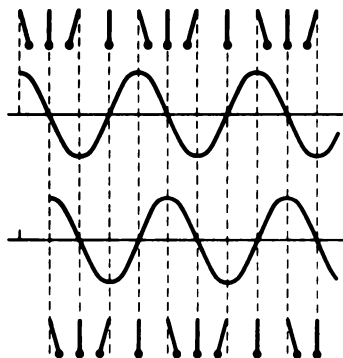


Fig. 11.

Oscillations of pendulums are shifted by a quarter of a period.

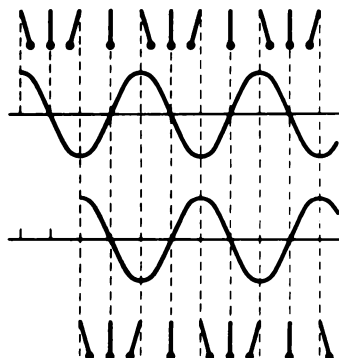
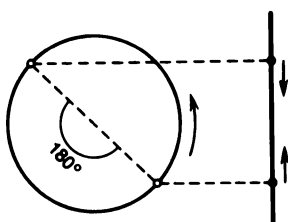


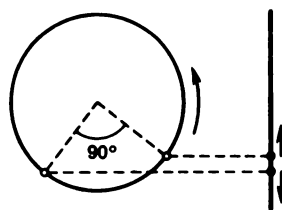
Fig. 12.

Oscillations of pendulums are shifted by half a period.





**Fig. 13.**  
Oscillations of shadows are  
shifted in phase by  $180^\circ$ .



**Fig. 14.**  
Oscillations of shadows are  
shifted in phase by  $90^\circ$ .

Such oscillations shifted in time can easily be obtained in the experiment with the shadow projections. If two balls are fixed at two diametrically opposite points of a uniformly rotating disc (Fig. 13), their shadows will oscillate with a shift of half a period, i.e. will always move in the opposite directions, simultaneously passing through the middle position. In order to obtain a shift of a quarter of a period, the balls should be arranged at a central angle of  $90^\circ$  to each other (Fig. 14). In this case, one shadow passes through the middle position when the other has the maximum deviation. In general, oscillations of shadows will be shifted by a part of a period equal to the ratio of the central angle between the radii, where the balls are fixed, to the complete circle ( $360^\circ$ ).

Oscillations of the same frequency shifted in time are said to have a *phase shift*. The time shift is expressed in fractions of a period, while the phase shift, or phase difference, is measured in angle units (degrees or radians).

If the second oscillation lags behind the first one by  $1/8$  of a period, this means that it lags in phase by  $360^\circ \times 1/8 = 45^\circ$ , or shifted in phase by  $-45^\circ$ . If the second oscillation leads the first one by  $1/8$  of a period, it is said to lead it in phase by  $45^\circ$ , or is shifted in phase by  $+45^\circ$ .

If there is no lag in two oscillations, they are termed *synphase* or are said to be *in phase* (i.e. in the same phase). If one oscillation lags behind the other one by half a period, the oscillations are said to be *in antiphase*.

The concept of phase shift, or phase difference, is obviously characterized by the time relation between two harmonic oscillations. However, we can speak about the phase of a single harmonic oscillation. *The phase of a harmonic oscillation* is the angle corresponding to the time elapsed from a certain arbitrarily chosen instant of time. Naturally, a period of oscillations, as before, corresponds to  $360^\circ$ .

Thus, the phase of oscillations depends on the instant taken as the time reference point. However, the phase difference for two oscillations does not depend on this arbitrary choice.

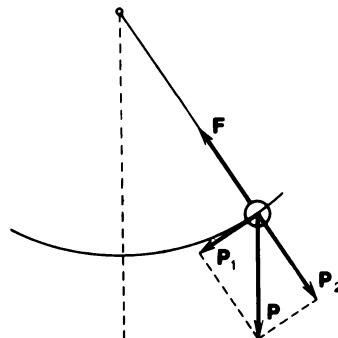
### 1.7. Dynamics of Pendulum Oscillations

Pendulums shown in Fig. 2 are elongated bodies of various shape and size, oscillating about the point of suspension or support. Such systems are known as *physical* (compound) *pendulums*. In equilibrium, when the centre of gravity lies on the vertical below the point of suspension (or support), the force of gravity is balanced (through elastic forces of the deformed pendulum) by the reaction of the support. When the pendulum is deflected from the equilibrium position, the force of gravity and elastic forces determine at each moment the angular acceleration of the pendulum, i.e. determine the nature of its rotation (oscillation). We shall now consider in detail the dynamics of oscillations by using as an example a *simple pendulum* consisting of a small load suspended on a long and thin thread.

In a simple pendulum, we can neglect the mass of the thread and the deformation of the load, i.e. assume that the mass of the pendulum is concentrated in the load, and the elastic forces are concentrated in the thread which is assumed to be unstretchable. Let us now analyze the forces which make the pendulum rotate after it has been drawn from the equilibrium position by this way or other (say, by a push or deflection).

When the pendulum is at rest in equilibrium position, the force of gravity acting downwards on the load is balanced by the tension of the thread. In a deflection position (Fig. 15), the force of gravity  $\mathbf{P}$  acts at an angle to the tension  $\mathbf{F}$  directed along the thread. We decompose the force of gravity into two components: along the thread ( $\mathbf{P}_2$ ) and at right angles to it ( $\mathbf{P}_1$ ). During oscillations of the pendulum, the tension  $\mathbf{F}$  slightly exceeds the component  $\mathbf{P}_2$  by the value of the centrifugal force that makes the load move along the arc of the circle. The component  $\mathbf{P}_1$  is always directed towards the equilibrium position and as if tends to restore equilibrium. For this reason, it is also referred to as the *restoring* force. The magnitude of  $\mathbf{P}_1$  is the larger, the greater the deviation of the pendulum.

Thus, when an oscillation (swinging) pendulum starts to deviate from the equilibrium position, say, to the right, a force  $\mathbf{P}_1$  emerges, which slows



**Fig. 15.**  
Restoring force  $\mathbf{P}_1$  acting on a  
pendulum in a deflected po-  
sition.

down its motion the more, the greater the deviation. Ultimately, this force brings the pendulum to a halt and drives it back to the equilibrium position. As this position is being approached, the force  $\mathbf{P}_1$  becomes smaller and smaller and vanishes when this position is attained. Thus, the pendulum passes through the equilibrium position by inertia. As soon as it starts deviating to the left, a force  $\mathbf{P}_1$  increasing with deviation will emerge again, but now it will be directed to the right. The motion to the left will be slowed down, and the pendulum comes to a halt for a moment and then will move with an acceleration to the right, and so on.

What energy transformations occur during oscillations of a pendulum?

The pendulum comes to a halt twice during a period, viz. when its deviations to the left and to the right are maximum. At these moments, its velocity is equal to zero, and hence its kinetic energy vanishes. But just at these moments the centre of gravity of the pendulum is elevated to the maximum height, and hence the potential energy attains its maximum value.

Conversely, when the pendulum passes through the equilibrium position, its potential energy attains the minimum value, while the velocity and the kinetic energy are at a maximum.

We shall assume that friction between the pendulum and air and the friction at the point of suspension can be neglected.<sup>3</sup> Then by the law of energy conservation, this maximum kinetic energy is exactly equal to the excess of the potential energy at the point of maximum deviation over the potential energy in the equilibrium position.

Thus, during pendulum oscillations, a periodic conversion of kinetic energy to potential energy and back takes place, the period of this process being equal to half the period of oscillations of the pendulum. However, the total energy of the pendulum (viz. the sum of its potential and kinetic energies) remains constant. It is equal to the energy imparted to the pendulum during its starting irrespective of whether it was the potential energy (initial deviation) or kinetic energy (initial impetus).

The same situation takes place for all oscillations in the absence of friction or some other processes taking away the energy from an oscillatory system or imparting an energy to it. For this reason, the amplitude remains unchanged and is determined by the initial deviation or the intensity of the impetus.

The same variations of the restoring force  $\mathbf{P}_1$  and the same energy conversion are observed if instead of suspending a ball on the string, we make it roll in the vertical plane in a spherical cup or in a circular trough. In this case, the role of the tension of the string will be played by the pressure exerted by the walls of the cup or trough (we again neglect the friction between the ball and air or walls).

---

<sup>3</sup> In Sec. 1.11, we shall take into account frictional forces.

### 1.8. Formula for the Period of a Simple Pendulum

The period of oscillations of a compound pendulum depends on many circumstances: the shape and size of a body, the distance between the centre of gravity and the point of suspension, and on the mass distribution in the body relative to this point. Therefore, it is rather difficult to calculate the period of oscillations of a suspended body. For a simple pendulum, the problem is much simpler. The following simple regularities can be established as a result of observation of simple pendulums.

1. If we suspend different loads on strings so that the length of the pendulums (viz. the distance between the point of suspension and the centre of gravity of the load) is the same, the period of oscillations will be the same although the masses of the loads differ considerably. *The period of oscillations of a simple pendulum does not depend on the mass of the load.*

2. If a pendulum is deflected through different (but not very large) angles and released, its period of oscillation will be the same, but the amplitudes will be different. As long as the amplitude is not very large, the oscillations are sufficiently close to harmonic oscillations (see Sec. 1.5), and the *period of a simple pendulum does not depend on the amplitude*. This property is known as *isochronism* (from the Greek words “isos” meaning equal and “chronos” meaning time).

This fact was established for the first time in 1655 by Galileo presumably under the following circumstances. Galileo observed oscillations of church-lustre in the Pisa cathedral. It was pushed during lighting and swang on a long chain. During the service, the swing of oscillations gradually damped (see Sec. 1.11), i.e. the oscillatory amplitude decreased, but the period remained unchanged. Galileo used his pulse for measuring time.

Let us now derive the formula for the period of oscillations of a simple pendulum.

As the pendulum swings, the load moves with an acceleration along the arc  $BA$  (Fig. 16a) under the action of the restoring force  $P_1$  which varies during motion. The calculation of the motion of a body under the action of a varying force is rather cumbersome. In order to simplify calculations, we shall proceed as follows.

We shall make the pendulum describe a cone instead of oscillating in the same plane, so that the load moves in a circle (Fig. 16b). This motion can be obtained as a result of superposition of two independent oscillations: one in the plane of the figure as before, and the other in the plane perpendicular to it. The periods of these plane oscillations are obviously equal since any plane of oscillations does not differ from any other plane in any respect. Consequently, the period of a composite motion, viz. the revolution of the pendulum along a cone, will be the same as the period of oscillations in one plane. This conclusion can be easily illustrated by

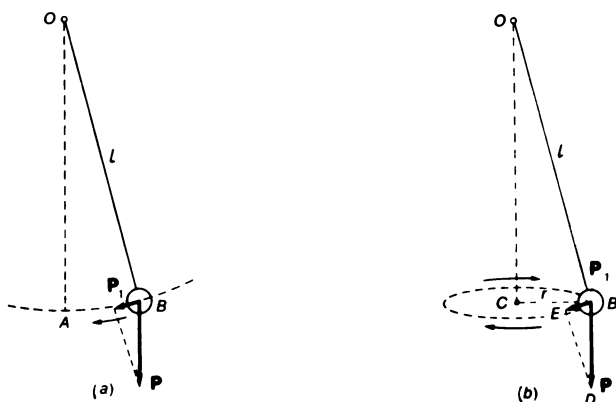


Fig. 16.

Oscillations of a pendulum in a plane (a) and its motion along a cone (b).

a direct experiment if we make one of two identical pendulums oscillate in a plane and the other revolve in a cone.

But the period of revolution of a “conical” pendulum is equal to the length of the circumference described by the load divided by the velocity:

$$T = \frac{2\pi r}{v}.$$

If the angle of deviation from the vertical is small (small amplitudes!), we can assume that the restoring force  $P_1$  is directed along the radius  $BC$  of the circle, i.e. is equal to the centripetal force:

$$P_1 = \frac{mv^2}{r}.$$

On the other hand, from the similarity of triangles  $OBC$  and  $DBE$  it follows that  $BE : BD = CB : OB$ . Since  $OB = l$ ,  $CB = r$ ,  $BE = P_1$  and  $BD = P = mg$ , we obtain

$$P_1 = \frac{r \cdot mg}{l}.$$

Equating two expressions for  $P_1$ , we obtain the following formula for the velocity of revolution:

$$v = r \sqrt{\frac{g}{l}}.$$

Finally, substituting this into the expression for the period  $T$ , we get

$$T = 2\pi \sqrt{\frac{l}{g}}.$$

Thus, the period of a simple pendulum depends only on the free-fall acceleration  $g$  and the length  $l$  of the pendulum, i.e. the distance between the point of suspension and the centre of gravity of the load. The above formula indicates that the period of a pendulum does not depend on its mass and amplitude (provided that the latter is sufficiently small). In other words, we have obtained as a result of calculations the basic regularities established earlier from observations.

However, the theoretical conclusion gives us more: it provides a quantitative dependence of the period of a pendulum on its length and free-fall acceleration. *The period of a simple pendulum is proportional to the square root of the ratio of the pendulum length to the free-fall acceleration.* The proportionality factor is equal to  $2\pi$ .

The dependence of the period of a pendulum on the free-fall acceleration forms the basis of a very accurate method for determining this acceleration. Having measured the length  $l$  of the pendulum and determined the period of oscillations  $T$  from a large number of oscillations, we can calculate  $g$  with the help of the formula for the period. This method is widely used in practice.

It is well known (see Vol. 1, Sec. 2.24) that the free-fall acceleration depends on the geographical latitude of a given point ( $g = 9.83 \text{ m/s}^2$  at the pole and  $g = 9.78 \text{ m/s}^2$  on the equator). The observations of the period of a standard pendulum make it possible to analyze the distribution of the free fall acceleration over latitude. This method is so accurate that it reveals even very small differences in the value of  $g$  on the surface of the Earth. It turns out that the values of  $g$  are different even at different points of the Earth's surface lying on the same latitude. These anomalies in the distribution of free-fall acceleration are associated with nonuniform density of the Earth's crust. They are used for analyzing the density distribution, in particular, for exploring minerals in the Earth's crust. Large-scale gravimetric measurements carried out in the USSR under the guidance of the Soviet physicist P.P. Lazarev revealed deposits of dense minerals in the region of the Kursk magnetic anomaly (see Vol. 2, Sec. 13.3). Combined with the data on the anomaly in the Earth's magnetic field, these gravimetric data were used for determining the distribution of iron ore deposits causing the Kursk magnetic and gravitational anomalies.

## 1.9. Elastic Vibrations

The restoring force acting on a simple pendulum is of gravitational origin. But for oscillations just the presence of a restoring force is essential, i.e. the force that is always directed to the equilibrium position and generally increases as the body moves away from this position. Forces of this type also appear during deformation of solids and are known as elastic forces (see Vol. 1, Sec. 2.29). Consequently, such elastic forces may also cause vibrations. Then vibrations are termed *elastic* according to the origin of restoring force. A few examples of this type of oscillation were considered earlier. The oscillations of a body suspended on a spring (such a device is often referred to as the simple oscillator), of a carriage on springs, of

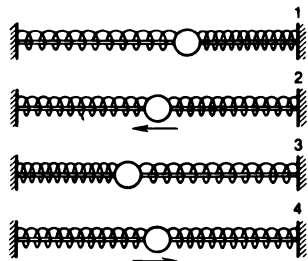
a plate gripped in a vice, of a tuning fork, of a stretched string, of a bridge, of a foundation, of a smoke chimney or of a high building are all examples of elastic vibrations.

A different origin of the restoring force is responsible for the fact that the potential energy of elastic vibrations is the energy of deformation of an elastic body and not the potential energy of the force of gravity as in the case of a simple pendulum. In all other respects, the dynamics of elastic vibrations is the same as that of a pendulum. Here too the kinetic energy is converted into the potential energy (of deformation) and back twice during a period.

All the stages of this process can be visually observed with a body (say, a ball) vibrating on two springs. In this case, we can assume that only the springs and not the ball possess the energy of deformation since the deformation of the ball can be neglected. If the mass of the ball is large in comparison with the mass of the springs, we can assume that only the ball possesses kinetic energy and not the springs whose mass is neglected. Thus, a transition from kinetic to potential energy and back is simultaneously an energy transfer from the ball to the springs and back.

Figure 17 shows four positions of this oscillatory system, which can be observed every quarter of a period. In position 1, the ball is in the extreme right position, the right-hand spring is compressed and the left-hand spring is stretched. The velocity and the kinetic energy are equal to zero, and the entire energy of the system is the potential energy. In position 2, the springs are not deformed, and the ball passes through the equilibrium position at the maximum velocity. The entire energy of the system is the kinetic energy of the ball. In position 3, the same situation as in position 1 is observed. Position 4 differs from position 2 only in the direction of velocity.

Taking a body with a larger mass and the same springs, we easily see that the vibrational frequency decreases. Using a stop-watch, we can verify that a four-fold increase in the mass of the body makes the period of vibrations two times longer (i.e. reduces their frequency to half the initial value). If the mass has been increased nine-fold, the period increases thrice. *The*



**Fig. 17.**  
Vibrations of a body on springs.

*period of elastic vibrations is proportional to the square root of the mass of the body.* This result is obtained in experiments with a degree of accuracy which is the higher, the closer we approach the above conditions, i.e. when we assume that the mass of the system is concentrated at a single point (the centre of mass of the body) and the mass of the springs is neglected. However, in all cases an *increase in the mass* of an elastic oscillatory system leads to *slowing down of vibrations*, i.e. *to an increase in their period.*

Let us now repeat the experiment with the body of the previous mass but with springs having a higher rigidity. It will be immediately seen that the period of vibrations decreases. Thus, *the period of elastic vibrations is the shorter, the higher the rigidity of the spring, i.e. the smaller the elasticity of the system.*

An analysis of elastic vibrations of a load on a spring indicates that for not very large amplitudes, these oscillations are of harmonic type, their period being expressed by a formula similar to the one for the period of a simple pendulum:

$$T = 2\pi \sqrt{\frac{m}{k}}.$$

Here  $m$  is the mass of vibration load and  $k$  is the rigidity of the spring, i.e. the force required to stretch the spring by a unit length.

### 1.10. Torsional Vibrations

An important particular case of elastic vibrations are *torsional vibrations* in which a body is twisted about its axis passing through its centre of gravity, first in one direction and then in the other.

If, for example, we suspend a disc on a wire (Fig. 18) and turn it so that the wire is twisted and then released, the disc will untwist the wire, then twist it by inertia in the opposite direction, and so on, i.e. will perform torsional vibrations. Here too the kinetic energy of the moving disc is transformed into the potential energy (of deformation) of the twisted wire and back twice during a period. Torsional vibrations often take place in the shafts of motors, for one,

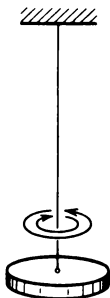


Fig. 18.

Torsional vibrations of a disc suspended on a wire.

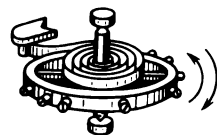
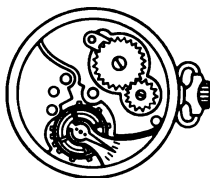


Fig. 19.

Balance wheel.



in screw shafts of steamers. Under certain conditions which will be specified below, torsional vibrations may turn out to be very harmful (Sec. 1.15).

In wrist and pocket watches, a suspended pendulum cannot be used, and a *balance wheel* (Fig. 19) is used instead, i.e. a wheel with a spiral spring ("hair spring") attached to its axle. The balance wheel vibrates so that the hair spring is bent (twisted or untwisted) in either direction from its equilibrium position. Thus, a balance wheel is a *torsional pendulum*.

The same regularities as for any elastic vibrations remain in force for torsional vibrations: the period is the longer, the smaller the rigidity of the system and the larger its mass (with the same shape).

In torsional vibrations, it is not only the mass that is important but also its distribution relative to the rotational axis. If, for example, we suspend on a wire a dumb-bell consisting of a knitting needle with two identical loads 1 and 2 fixed symmetrically to it (Fig. 20), as we move the loads apart, the frequency of torsional vibrations will decrease although the mass of the dumb-bell remains unchanged. Taking heavier loads instead of 1 and 2 and preserving the same separation between them, we see that the frequency also becomes lower.

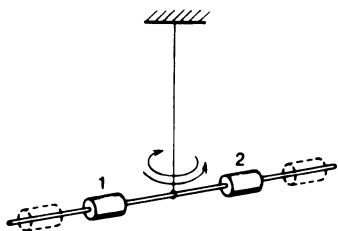


Fig. 20.  
Torsional vibrations of a dumb-bell.

With small torsion angles (small angular amplitudes), torsional vibrations are also of the harmonic type. Their period is determined by the relation

$$T = 2\pi \sqrt{\frac{I}{k}}.$$

where  $k$  is the rigidity of the system. The rigidity  $k$  is numerically equal to the torque required for the rotation through a radian. If the elastic forces are due to the twisting of the wire of string, then  $k$  is the so-called torsional rigidity of these bodies. The quantity  $I$  characterizes the mass distribution relative to the rotational axis (it is the so-called *moment of inertia* that plays the same role in rotational motion as the mass does in translatory motion). For example,  $I = 2mr^2$  for a dumb-bell, where  $m$  is the mass of each load and  $r$  is the distance from the loads to the rotational axis.

### 1.11. Effect of Friction. Damping

While considering free oscillations of a pendulum, a ball with springs or a disc, we ignored so far the phenomenon which always occurs in each of the experiments described above and due to which oscillations are not strictly periodic. Namely, the amplitude of oscillations becomes smaller and smaller with each swing so that ultimately oscillations attenuate. This phenomenon is known as the *damping of oscillations*.

The reason behind damping consists in that besides the restoring force, various types of friction that hamper motion always act in any oscillatory system (air resistance and so on). During each swing, a fraction of the total (potential and kinetic) energy is spent to do work against friction. Ultimately, the entire energy initially imparted to the oscillatory system and stored in it is converted into this work (see Vol. 1, Secs. 4.17-4.19).

The energy may be spent to do work against friction in various ways. There can be friction between solid surfaces, say, between the fulcrum of the beam of a balance and the support. The energy can be spent to overcome the resistance of the medium (air or water) (see Vol. 1, Secs. 2.35, 2.36). Besides, vibrating bodies set in motion the surrounding medium, giving away a fraction of their energy in each oscillation (see Vol. 1, Sec. 2.38). Finally, the deformations of springs, plates, wires, etc. also proceed with a certain loss of energy associated with internal friction in the material from which a body is made (see Vol. 1, Sec. 11.1).

*Free undamped oscillations occurring in an oscillatory system in the absence of friction* are called *natural oscillations* of the system.

Disregarding frictional forces so far, we hence considered just these idealized strictly periodic natural oscillations, intentionally simplifying an analysis of oscillations at the expense of a slightly inaccurate description. This simplification, however, is possible and applicable since friction and damping caused by it are actually small in many oscillatory systems: a system has time to perform a large number of oscillations before their amplitude becomes noticeably smaller. While studying such systems with low damping, this damping can be completely ignored in solving a large number of problems, and free oscillations of a system can be assumed to be strictly periodic. In other words, we can consider natural oscillations as it was done above.

An oscillation which would be harmonic in the absence of damping (natural oscillation) is of course no longer harmonic in the presence of damping. In addition, the motion will not be periodic any longer due to damping. Its oscillogram is not a curve repeating itself (Fig. 21) but rather a curve with a decreasing amplitude (Fig. 22). Increasing friction in one way or another, we can obtain so strong damping that the system comes

Fig. 21.  
Undamped oscillations.

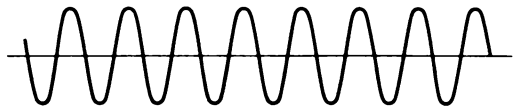


Fig. 22.  
Damped oscillations.



to a halt after the very first swing or even before it goes over through the equilibrium position (Fig. 23). Such strongly damped systems are termed aperiodic.

Using vibrations of a load on a spring, we can easily observe an increase of damping with friction. If we place the load in water, the damping of vibrations sharply increases as compared with damping in air. In oil, the damping will be still stronger than in water: we obtain an aperiodic or nearly aperiodic motion. The less streamlined is the shape of the load (for the same mass), the stronger the damping since the larger is the fraction of energy spent for setting in motion the surrounding medium (see Vol. 1, Sec. 9.11).

In practice, damping has to be reduced in some cases or enhanced in other. For example, the axle of the balance wheel of a watch has pointed ends which rest on highly polished conic supports made of hard stone (agate or ruby) called jewels. This is done to reduce the damping of the balance wheel. On the contrary, while taking measurements with many measuring instruments, it is desirable that the movable part of an instrument stop soon without performing a large number of oscillations, or even aperiodically. For this purpose, various types of *dampers* are used, viz. the devices that increase friction and the energy loss in general. A damper can be made in the form of a plate immersed in oil. Electromagnetic dampers (Fig. 24) are also employed, where a metallic plate moving between the poles of an electromagnet is slowed down due to eddy electric currents induced in it (see Vol. 2, Sec. 15.6.).

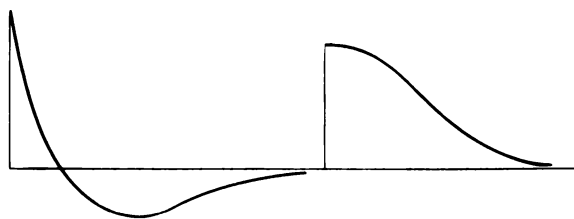


Fig. 23.  
Aperiodic motion.

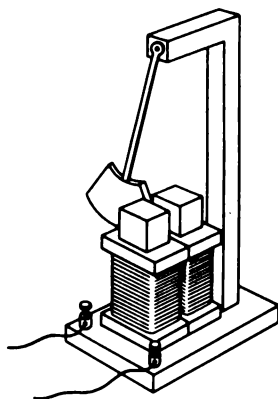


Fig. 24.  
Pendulum damped by eddy currents.

Friction affects not only the amplitude of oscillations (damping) but also the duration of swings. This duration cannot be called a period since a damping oscillation is an aperiodic motion. If, however, the damping is not strong, we can conventionally speak about a period, having in mind the time elapsed between two passages through the equilibrium position in the same direction. The period becomes *longer with increasing friction*.

A typical feature of oscillatory systems is that the effect of a small friction on the period of oscillations is much weaker than the effect on the amplitude. This circumstance has played a very important role in the perfection of clocks. The idea of using a pendulum, viz. an oscillatory system in clocks belongs to Galileo.

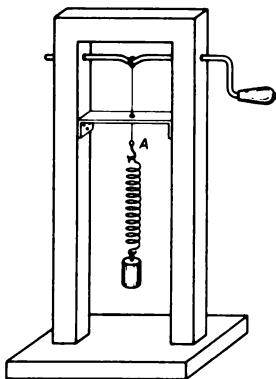
The first pendulum clock was constructed in 1673 by the Dutch physicist and mathematician Ch. Huygens (1629-1695). That year can be considered as the date of birth of modern clocks which pushed into oblivion all clockwork mechanisms that existed before. This mainly occurred due to the fact that the pace of a clock with the pendulum is not sensitive to the variation of friction that depends on so many factors. In previous clocks without pendulum (like water clock, see Vol. 1, Sec. 1.8), the pace strongly depended on friction.

### 1.12. Forced Vibrations

So far, we considered free oscillations, i.e. periodic motions performed by an oscillatory system disturbed from equilibrium and left to itself. It was mentioned above, however, that in some cases a periodic motion is not free but occurs under the action of a periodically varying force. Such repeatedly acting forces cause periodic motion of bodies which themselves are not oscillatory systems. Let us recall, for example, a periodically opened and closed door or the motion of the needle in a sewing machine. It can be easily seen that the period of motion caused by a periodically varying force is equal to the period of the force.

But what will be if a periodic force acts on an oscillatory system? We know that any oscillatory system has its own period, i.e. the period of natural oscillations, and the force may vary with some other period. What will then be the period of motion?

Let us take for an oscillatory system a load suspended on a spring, and suspend this simple oscillator at the end of the string of the device described in Sec. 1.5 (Fig. 25). If we uniformly rotate the handle, the motion of the point of suspension of the simple oscillator allows us to exert a harmonic force on the oscillator. The period of variation of this force is obviously equal to the period of rotation of the handle.



**Fig. 25.**

Forced vibrations of a load on a spring.

When we start to rotate the handle uniformly, the load is set in motion which at first appears rather complex. However, after completing a few turns we shall see that the motion of the load has become a regular periodic vibration. Here irrespective of the speed of rotation of the handle, the steady-state vibrations of the simple oscillator have a period equal to the period of rotation of the handle. Thus,

(1) *in an oscillatory system subjected to a periodically varying force, a periodic motion sets in*; in contrast to natural vibrations, the periodic motions of this type are termed *forced vibrations*;

(2) *the period of forced vibrations is equal to the period of acting force*.

It was shown above that free vibrations are damped due to friction and other losses of energy. They are undamped only in the ideal case when friction is absent completely (natural vibrations). Forced vibrations are, in spite of friction, actually periodic and repeating themselves as long as the periodic force acts on the system. This is due to the fact that in forced vibrations the energy spent for friction is continually replenished at the expense of the work done by the periodic force acting on the system, while in free vibrations an energy is supplied to the system only at the beginning of motion, and the motion continues until this store of energy is exhausted.

### 1.13. Resonance

Let us now consider in greater detail the amplitude of the forced vibrations of the load in the experiment described in the previous section. At different speeds of rotation of the handle, i.e. at different periods of coercive force, the amplitude is far from being the same.

If we rotate the handle slowly, say, completing one revolution during 3-5 s, the load and the spring will move up and down in the same way as the point *A* (Fig. 25). Thus, the amplitude of forced vibrations of the load will be the same as the amplitude of the point *A*. For a more rapid rotation, the load starts vibrating at a larger amplitude. The amplitude of forced vibrations becomes very large (a few times larger than the amplitude of the point *A*) if the period of rotation of the handle, i.e. the period of the force, is made close to the period of natural vibrations of the load on the spring. If, however, we shall rotate the handle still more rapidly, the amplitude of forced vibrations decreases again. Finally, a very rapid rotation of the handle leaves the load almost at rest.

The same can be observed for forced vibrations of a simple pendulum. The periodic force in this case can be simply created by rocking the holder from which the pendulum is suspended.

These and many other similar experiments show that when a periodic force acts on an oscillatory system, the case when the period of variation

of the force coincides with the period of free vibrations of the system is of special importance.

The coincidence of the period of free vibrations of a system with the period of an external force acting on this system is called the *resonance*. Thus, the *amplitude of forced vibrations attains the maximum value at resonance*.

Phenomena emerging during resonance (like the attainment of the maximum amplitude of forced vibrations) are known as *resonance phenomena*. A force whose period coincides with the period of free vibrations and which thus causes the build-up of vibrations, i.e. the maximum “response” of the oscillatory system is said to act in resonance. A system whose period is equal to the period of the external force is said to be tuned to resonance. Naturally, if damping is weak so that the period of free vibrations is close to the natural vibrations period (see Sec. 1.11), the resonance tuning can be interpreted as the coincidence of the period of the coercive force with the period of natural vibrations.

The phenomena of resonance can be demonstrated with the help of the following simple experiment. A heavy pendulum 1 and light pendulums having different lengths (and hence different periods) are suspended from the same holder (Fig. 26).

We make pendulum 1 swing in the plane perpendicular to the plane of the holder. Its vibrations will cause a periodic bending of the holder bar so that a force will act on all the remaining pendulums with the period of pendulum 1. It will be seen that pendulums 2 and 3 whose periods differ from that of pendulum 1 stronger almost remain at rest, i.e. their amplitudes are very small. Pendulums 4 and 5 whose periods are closer to that of pendulum 1 will oscillate with a larger amplitude. Finally, pendulums 6 and 7 having the same length as pendulum 1, i.e. tuned to resonance, swing with a very large amplitude.

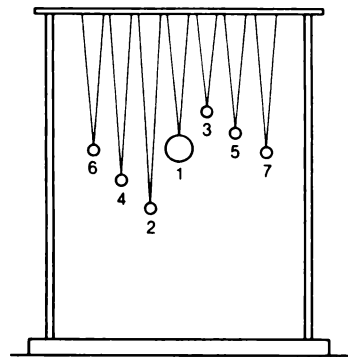


Fig. 26.

Demonstration of resonance by using pendulums.

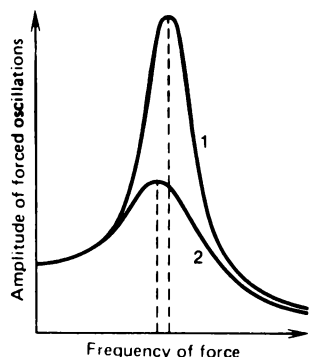
### 1.14. Effect of Friction on Resonance Phenomena

Friction which is present in a system and responsible for the attenuation of its free vibrations is very important for resonance phenomena. This can be easily demonstrated in experiments on vibrations of a simple oscillator (Fig. 25) with varied friction. Damping can be increased by immersing the load, as before, in water or oil.

When the load vibrates in air, the coincidence of the period of rotation of the handle of the device (i.e. the period of force) with the natural vibration period of the system leads to a strong build-up, i.e. the amplitude of vibrations of the load is several times larger than the amplitude of the point *A* of the device. But a slight speeding up or slowing down of the rotation of the handle is sufficient for the amplitude of the load to be abruptly reduced. Thus, *if the damping of a system is weak, the resonance phenomena are manifested strongly and clearly (sharp resonance)*. The build-up for an exact resonance is very large, but a small detuning (difference between the period of the force and the period of natural vibrations of the system) considerably reduces the amplitude of forced vibrations.

On the contrary, for a damped system, i.e. the system with increased damping (e.g., the load vibrating in water), the amplitude of forced vibrations at an exact resonance does not considerably exceed the amplitude of vibrations of the point *A*. On the other hand, at a deviation from resonance to either side, the amplitude decreases not so sharply. For example, if we increase the speed of rotation of the handle twice as compared to the resonance frequency, the vibrations of the load immersed in water will turn out to be only slightly smaller than in resonance. On the contrary, vibrations of the same load in air will be much weaker. Thus, *if damping is strong, the resonance phenomena are weak and not clearly manifested (broad resonance)*: an increase in amplitude at exact resonance is relatively small, and the amplitude falls down noticeably only at a large detuning.

These results are illustrated by the graph shown in Fig. 27. The figure represents so-called *resonance curves* describing the dependence of the am-



**Fig. 27.**  
Resonance curves for weak damping (1) and strong damping (2).

plitude of forced oscillations on their frequency, i.e. the frequency of the force acting on the system. Two curves corresponding to weak and strong damping are shown in the figure. The first curve has a narrow and large maximum, while the second curve has a small and broad maximum. Pay attention to the fact that the first curve is higher than the second curve everywhere, i.e. at any frequency of the external force the amplitude of forced vibrations is the higher, the weaker the damping. At exact resonance, the difference in the amplitudes of vibrations with weak and strong damping is especially large.

Besides, the maximum of curve 2 is slightly shifted to the left of the maximum of curve 1, i.e. corresponds to a slightly lower frequency of the external force. This is due to an increase in the period of free vibrations at increased damping.

It should be borne in mind that resonance curves give the values of the steady-state amplitude. Vibrations with this amplitude set in not immediately but during a certain time, starting from the moment when the force began to act on the system. This was pointed out in Sec. 1.12 when we described the emergence of forced vibrations of the load on the spring.

How long does the stabilization period last?

This question can easily be answered if we take into account the fact that at the initial moment, when a periodic force starts acting on the system, free vibrations emerge in it along with forced vibrations. It is for this reason that the initial motion of the system is complex since it is the superposition of two motions: forced vibrations with the frequency of the force and free vibrations with the natural frequency. But the force sustains only forced vibrations, while free vibrations attenuate, and hence the motion is gradually "cleared off" from them. Only forced vibrations remain in the system.

Thus, *the process of stabilization of forced vibrations consists in that free oscillations, induced at the moment when the force starts acting and mixed with forced vibrations, attenuate.* For this reason, the process of stabilization of forced vibrations takes the same time as the process of damping of free vibrations. But this means that *with a very weak damping, the resonance amplitude is very large, but the build-up to this amplitude takes a long time. Conversely, with a strong damping of the system, the resonance amplitude is not large, but it stabilizes rapidly.* This should be taken into account while making the experiments described above.

### 1.15. Examples of Resonance Phenomena

Resonance plays a very important role in various phenomena, being useful in some of them and harmful in others. Let us consider a few examples involving mechanical vibrations.

Crossing a ditch with the help of a board thrown across it, one may get in resonance with the natural vibration period of the system (including the board and the person on it). The board starts to vibrate vigorously (bending upwards and downwards). The same may happen to a bridge crossed by a squadron of infantry or by a train (the periodic force emerges



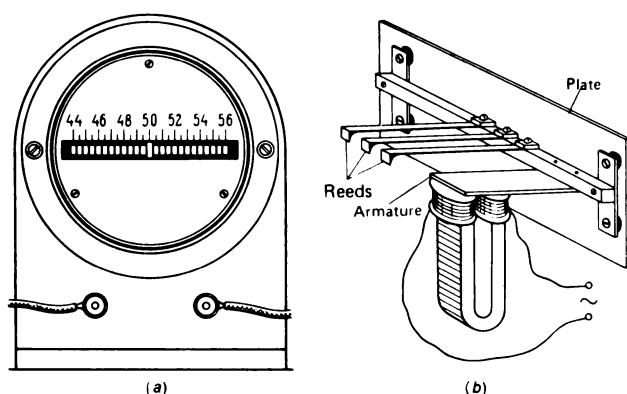
as a result of marching in unison or due to impacts of the wheels at the joints between the rails). For example, in 1906 the Egyptian bridge across the Fontanka river in St. Petersburg broke down when a cavalry squadron was passing through it. The measured pace of well-trained horses got in resonance with the period of natural vibrations of the bridge. In order to prevent such events, troops are ordered “to break step” while marching over a bridge. Trains normally pass over bridges slowly to make the period of impacts of the wheels against the joints of rails much longer than the period of natural vibrations of a bridge. Sometimes the inverse method of “detuning” is used, and a train hurtles over a bridge at the maximum velocity.

It may happen that the period of impacts of the wheel against the joints of rails coincides with the period of vibrations of a carriage on shock absorbers. Then the carriage swings heavily. A ship also has its own period of rocks. If sea waves get in resonance with the ship movements, the rocking becomes vigorous. In this case, the captain orders to change the velocity of the ship or its course. As a result, the period of waves running onto the ship is changed (due to the change in the relative velocity of the ship and waves), and the system is detuned.

Unbalancing of machines and motors (eccentricity, or bending of the shaft) gives rise to a periodic force acting on the support of a machine (foundation, hull of the ship, etc.) during its operation. The period of this force may coincide with the period of free vibrations of the support or, for example, with the period of bending vibrations or torsional vibrations of the rotating shaft itself. Then the resonance is observed, and forced vibrations may become so strong that foundations are destroyed, shafts break, and so on. In such cases special measures should be taken to avoid resonance or weaken its effect (detuning of periods, damping, and so on).

Obviously, in order to obtain a certain amplitude of forced vibrations with the help of the minimal periodic force, resonance conditions must be observed. The heavy reed of a bell may be swung even by a child if he stretches the rope with a period of free oscillations of the reed. However, even a very strong person cannot do that by pulling the rope out of resonance.

Resonance underlies the operating principle of an instrument for measuring the frequency of alternating current varying according to a harmonic law (see Vol. 2, Sec. 17.3). Such instruments are known as *vibrating reed frequency meters* and are used for controlling the constancy of frequency in electric circuits. The general view of the instrument is shown in Fig. 28a. It consists of a set of elastic plates with loads (reeds) at the ends. The masses of the loads and the rigidities of plates are chosen so that the frequencies of neighbouring reeds differ by the same number of hertz. In the frequency meter shown in Fig. 28a, the frequencies of the reeds differ by 0.5 Hz.

**Fig. 28.**

Vibrating reed frequency meter: (a) general view and (b) schematic diagram.

These frequencies are indicated on the scale against corresponding reeds.

The schematic diagram of this instrument is shown in Fig. 28b. The current under investigation is passed through the winding of the electromagnet. Vibrations of the armature are transferred to the plate fixed to flexible strips and connected to the free ends of the flexible plates with reeds. Thus, each reed experiences the action of a harmonic force whose frequency is equal to the frequency of the current. A reed which is in resonance with this force vibrates with the maximum amplitude and indicates its frequency, and hence the frequency of the current, on the scale.

Henceforth we shall encounter resonance phenomena more than once while studying acoustic and electric oscillations. The most striking examples of useful application of resonance will be provided by these types of oscillation.

### 1.16. Resonance Phenomena Induced by an Anharmonic Periodic Force

In the experiments described in Secs. 1.12-1.14, periodic actions were exerted by the bodies performing harmonic oscillations (the motion of the string in the device shown in Fig. 25 or a massive pendulum). Accordingly, the external force varied by a harmonic law. The observation that a sharp increase in amplitude takes place only when the period of the force coincides with the period of natural oscillations of the system refers just to this case.

What will be the effect of the force varying periodically according to a law differing from a harmonic law?

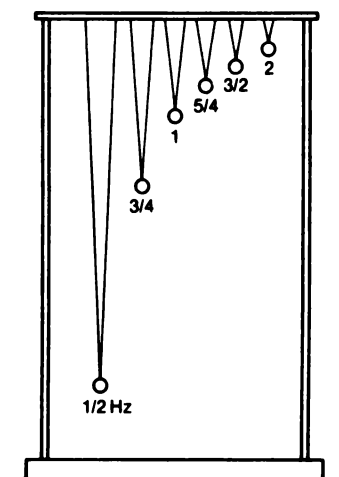
For example, we can periodically push a pendulum, i.e. act on it with short repeating impacts. Experiments show that in this case resonance phenomena are observed not only for a single period of the force. The

build-up of oscillations will take place as before as we push the pendulum once during a period of its free oscillations. But a noticeable build-up will also be observed if we push the pendulum less frequently, missing one, two, etc. swings.

Thus, it follows from the experiment described above that *if a force varies periodically but not according to a harmonic law, it may cause resonance phenomena not only when its period coincides with the period of free oscillations of a system, but also when the period of the force is longer than the oscillation period an integral number of times.*

The same conclusion can be drawn from the following experiment. Instead of one oscillatory system (pendulum) acted upon in turn by the forces having different periods, we can take a set of similar systems with different natural frequencies and act simultaneously with the same periodic force. In order to obtain a sharp resonance, the systems must have low damping. We shall use again a set of pendulums but not the same as in Fig. 26. The lengths of the largest and smallest pendulums differed in that experiment only by a factor of 2, i.e. the natural frequencies differed by a factor of  $\sqrt{2} = 1.4$ . Now we take pendulums with natural frequencies lying in a wider range and such that there are pendulums with multiple frequencies. Let, for example, the natural frequencies of the pendulums be  $1/2$ ,  $3/4$ , 1,  $5/4$ ,  $3/2$  and 2 Hz. The corresponding lengths of the pendulums will be approximately equal to 100, 44.4, 25, 16, 11.1 and 6.3 cm. This set of pendulums is shown in Fig. 29.

Naturally, we can verify that when a harmonic force acts, the pendulum which is tuned in resonance with the frequency of the force will be the only one that has the maximum amplitude.



**Fig. 29.**  
Set of pendulums whose frequencies are indicated in the figure.

A harmonic force can be obtained, as before, by suspending a massive pendulum from the same holder and making its length equal to the length of a pendulum from our set. Experiment can be also successfully made if we just swing the entire holder by hand, setting it in harmonic oscillations in unison with oscillations of one of the pendulums. Just this pendulum will swing with the maximum amplitude, the remaining pendulums being practically at rest.

The situation is quite different if instead of harmonic shaking of the holder we apply to it sharp periodic impacts, i.e. act on the pendulum with a periodic but anharmonic force. Pushing the holder with the period of the longer pendulum, viz. once in 2 s, we shall see that not only this pendulum but also some other pendulums oscillate. Namely, these will be the pendulums whose natural frequencies are higher than the frequency of the longest pendulum (1/2 Hz) an integral number of times. In other words, besides the pendulum whose frequency is equal to 1/2 Hz, the pendulums having the frequencies of 1, 3/2 and 2 Hz will oscillate vigorously, while the remaining pendulums will be practically at rest. Comparing this result with the previous case when the harmonic force induced oscillations on only one pendulum, we arrive at the following conclusion.

*Anharmonic periodic action with a period  $T$  is equivalent to a simultaneous action of harmonic forces with different frequencies, namely, with frequencies that are multiple to the lowest frequency  $\nu = 1/T$ .*

This conclusion about a periodic force is a special case of the mathematical theorem proved in 1822 by the French mathematician J.B. Fourier (1768-1830). The Fourier theorem states that *any periodic oscillation of period  $T$  can be represented in the form of the sum of harmonic oscillations with periods equal to  $T$ ,  $T/2$ ,  $T/3$ ,  $T/4$ , . . . , i.e. with frequencies  $\nu = 1/T$ ,  $2\nu$ ,  $3\nu$ ,  $4\nu$ , . . .*

The lowest frequency  $\nu$  is known as the *fundamental frequency*. An oscillation with the fundamental frequency  $\nu$  is termed the *first harmonic*, or *fundamental tone* while oscillations at frequencies  $2\nu$ ,  $3\nu$ ,  $4\nu$ , etc. are called *higher* (second, third, fourth, etc.) *harmonics* or *overtones*.

The Fourier theorem is a mathematical theorem of quite general nature which makes it possible to represent any periodic quantity (displacement, velocity, force, etc.) as a sum of quantities (displacements, velocities, forces, etc.) varying according to the sine law.

When applied to the problem under consideration on the effect of a periodic anharmonic force, this theorem immediately explains why a pendulum can be swung not only as a result of pushing it successively with a period equal to its own period but also with twice, thrice, etc. longer period.

Let the natural frequency of a pendulum be 1 Hz. Pushing it once a second, we create a periodic force comprising the following harmonic oscil-

lations: the fundamental frequency of 1 Hz and overtones with frequencies of 2, 3, 4, etc. Hz. Thus, in this case the first harmonic of the force is in resonance with the natural frequency of the pendulum. If we push the pendulum with a period twice as long, viz. once in 2 s, the force will be the sum of oscillations with the fundamental frequency of 1/2 Hz and with harmonics of 1, 3/2, 2, 5/2, etc. Hz.

Consequently, the amplitude of the pendulum increases since the second harmonic of the force is in resonance with the natural frequency of the pendulum. If we repeat pushes in 3 s, the third harmonic of the force will be in resonance with the natural frequency of the pendulum, and so on.

Thus, *a periodic anharmonic force strongly increases the amplitude of an oscillatory system if one of harmonic oscillations constituting the force is in resonance with the natural frequency of the system.*

The reed frequency meter described in Sec. 1.15 can be used in the same way as a set of similar pendulums for the harmonic analysis of a unharmonic force mentioned at the beginning of this section.

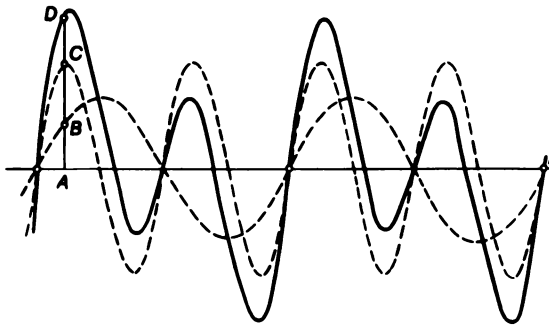
It was shown above that one of the reeds of the frequency meter starts oscillating under the action of a harmonic force. For any unharmonic effect (say, intermittent current), not one but all those reeds which are in resonance with harmonics constituting the current will oscillate. The amplitude of each reed in this case is directly proportional to the amplitude of the harmonic component of the current, which is in resonance with the natural frequency of the reed. Frequency meters can also be used for determining the harmonic composition of mechanical vibrations, for example, the vibrations of the foundation of a machine. For this it is sufficient to place the instrument onto the vibrating foundation.

### 1.17. The Relation Between the Form and Harmonic Composition of Periodic Oscillations

We can now answer the question posed in Sec. 1.5: what does it mean that an anharmonic periodic oscillation of period  $T$  has no definite frequency?

According to the Fourier theorem, such a periodic oscillations is a set of harmonic oscillations and is hence characterized by a set of frequencies  $\nu = 1/T, 2\nu, 3\nu, \dots$ , i.e. the multiples of the lowest (fundamental) frequency  $\nu$ , rather than by a single frequency.

Let us consider the oscillograms of oscillations having the same period  $T$  but different forms. An example of such oscillograms was represented in Fig. 6 showing a few different periodic oscillations having the same period. According to the Fourier theorem, each such oscillation is the sum of harmonic oscillations, so that the fundamental frequency  $\nu = 1/T$  and higher harmonics  $2\nu, 3\nu, \dots$ , are the same for all periodic oscillations under consideration, since they have the same period  $T$ .

**Fig. 30.**

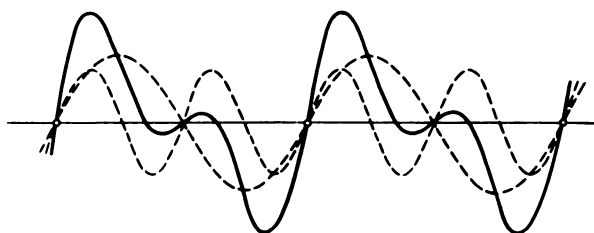
The sum of a harmonic oscillation and its second harmonic.

But if the frequencies of harmonics are the same, then why do the periodic oscillations have different forms?

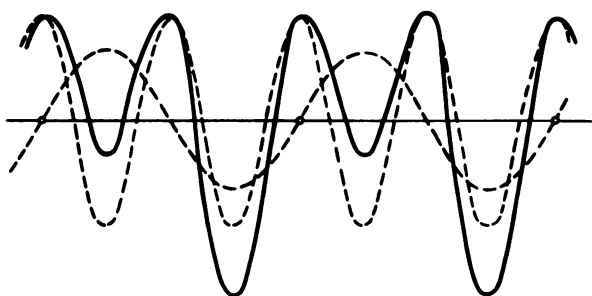
Let us try to answer this question by analyzing the summation of harmonic oscillations. This summation is carried out according to the rules of summation of motions (see Vol. 1, Sec. 1.6). If the displacements being added take place along the same straight line, the resultant displacement is equal to the algebraic sum of the displacements being added. Thus we obtain the graphical method of summation of oscillations, which we are going to use.

The dashed lines in Fig. 30 represent the oscillograms of two harmonic oscillations, viz. the fundamental (first) and second harmonics. The horizontal axis corresponds to the equilibrium position. At a certain instant of time, i.e. at a certain point *A* of this straight line, we have segments *AB* and *AC* representing the deviation from the equilibrium position caused by each oscillation at this instant. Summing up these segments, we obtain segment *AD* representing the resultant deviation at point *A*. Proceeding in the same way for a number of points on the straight line and considering that the deviations have the plus sign if they lie above the horizontal axis and the minus sign if they lie below it, we connect the ends of the resultant segments by a curve. This is the scan of the resultant oscillation (solid curve in the figure); it has the same period as the first harmonic but its form differs from the sinusoid.

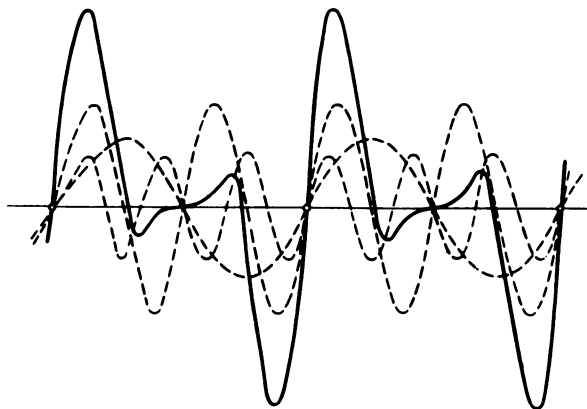
Let us now reduce the amplitude of the second harmonic by half. The result of summation in this case is shown in Fig. 31. In Fig. 32, the amplitudes of both harmonics are the same as in Fig. 30, but the second harmonic is shifted in time by a quarter of its period. Finally, in Fig. 33 the harmonics are the same as in Fig. 30, but one more (the third) harmonic is added. In all cases, the resultant oscillations have the same period but differ in the form considerably.

**Fig. 31.**

Same as in Fig. 30, but the amplitude of the second harmonic is reduced by half.

**Fig. 32.**

Same as in Fig. 30, but the second harmonic is shifted by a quarter of its period.

**Fig. 33.**

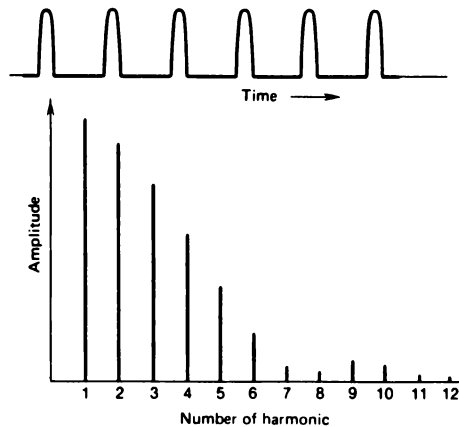
Same as in Fig. 30, but the third harmonic is added.

Thus, the *difference in the form of periodic oscillations is determined by the number of harmonics in their composition as well as by their amplitudes and phases.*

For the sake of simplicity, we added only two or three harmonics. But the forms of periodic oscillations can be (and often are) such that the number of higher harmonics is very large or even infinitely large. For each form of a periodic oscillation, every harmonic constituting it has a certain amplitude and phase. As soon as the amplitude or phase of a single harmonic has changed the form of the resultant periodic oscillation changes to a certain extent.

However, the changes in the form of oscillations due to the phases of harmonics, i.e. due to their shifts in time, are often unimportant for a physical phenomena and hence are not taken into account. This is the case, for example, with acoustic vibrations which will be discussed in the following sections. Here, it is important to know just the frequencies and amplitudes of harmonics constituting a given complex oscillation. The set of such frequencies and amplitudes is termed the *harmonic spectrum* (or just *spectrum*) of a given oscillation.

Spectra can be represented in the form of very illustrative graphs by plotting the frequencies (or numbers) of harmonics on a certain scale along the horizontal axis, and their amplitudes along the vertical axis. Figure 34 represents the oscillogram of an oscillation in the form of periodic spikes in one direction. This corresponds, for example, to the time variation of a pulsed force. The lower part of the figure represents the spectrum of this oscillation. The position of every line is determined by the number of the corresponding harmonic, and hence its frequency, while the height of the line represents the amplitude of this harmonic.



**Fig. 34.**

Periodic oscillation in the form of spikes and its spectrum.



## Chapter 2

# Acoustic Vibrations

### 2.1. Acoustic Vibrations

An elastic plate fixed in a vice vibrates at a frequency which is the higher the shorter the free (vibrating) end of the plate. When the frequency of vibrations becomes higher than 16 Hz, we can hear the vibrations of the plate. In Sec. 1.4 it was shown that a sounding tuning fork also vibrates. In general, a human ear perceives a sound when it is affected by mechanical vibrations at a frequency not lower than 16 Hz and not higher than 20 000 Hz (20 kHz).<sup>1</sup> Vibrations at higher or lower frequencies cannot be heard.

Thus, *sounds are caused by mechanical vibrations in elastic media or bodies (solid, liquid or gaseous), with frequencies lying in the range from 16 to 20 000 Hz and which can be perceived by a human ear.*

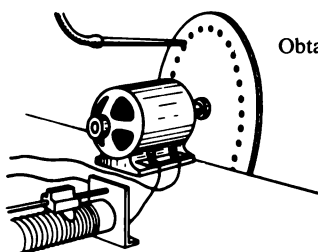
Accordingly, mechanical vibrations at frequencies indicated above are called *sonic*, or *acoustic* vibrations (acoustics is the branch of physics dealing with sounds). Inaudible mechanical vibrations at frequencies below the acoustic range are often termed *infrasonic* vibrations, while those at frequencies higher than the sonic range, i.e. above 20 kHz, are known as *ultrasonic* vibrations.

If we place a sounding body, say, an electric bell, under the bell of an air pump, the sound becomes weaker and weaker as the air is being pumped out, and finally becomes inaudible. Oscillations propagate from a sounding body through air. The process of propagation of oscillations in air will be considered later. At the moment, we shall note only one circumstance: a sounding body alternately compresses the layer of air adjoining its surface or creates a rarefaction in this layer during its vibrations. Thus, the propagation of sound in air starts with oscillations of the air density near the surface of the vibrating body.

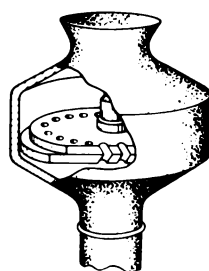
However, air density oscillations can be produced without a vibrating body. If, for example, we rapidly rotate a disc with holes arranged along the circumference and blow an air jet through it (Fig. 35), the jet becomes

---

<sup>1</sup> We mean that vibrations are inaudible themselves, i.e. if they are not accompanied by vibrations of other origin with audible frequencies. The squeak of swings does not mean that we can hear their oscillations.



**Fig. 35.**  
Obtaining sound by interrupting  
an air jet.



**Fig. 36.**  
Siren.

intermittent behind the holes, and periodic compressions of air will be produced. It can be easily verified that we shall hear a sound in this case.

The construction of a siren is just based on the interruption of an air jet. In this source of sound, a rotating disc is usually arranged above a stationary disc with the same number of holes drilled at an acute angle (Fig. 36). Firstly, this is done to rotate the movable disc by the air jet like the turbine wheel, and secondly, the number of interrupted jets is equal to the number of holes in the disc, due to which the sound is amplified considerably.

A siren or even a simple device shown in Fig. 35 are convenient for experiments since they allow one to easily determine the period of acoustic vibrations. The number of interruptions of an air jet per second is obviously equal to the product of the number  $z$  of the holes and the number  $n$  of revolutions of the disc per second. The period is the reciprocal to this quantity:

$$T = \frac{1}{zn}.$$

Since air vibrations caused by a siren are anharmonic, the number of interruptions ( $zn$ ) of the air jet is not the frequency of vibrations. As was mentioned above, a periodic anharmonic motion cannot be characterized by a single frequency but rather contains a set of harmonic oscillations with frequencies multiple to the fundamental frequency  $\nu = 1/T$  (see Sec. 1.17).

## 2.2. Subject of Acoustics

Problems studied in acoustics are quite diverse. Some of them are ultimately related with the properties and peculiarities of our hearing. Physiological acoustics investigates the hearing organ (ear), its construction and operation. Architectural acoustics studies the propagation of sound in rooms, the effects of the shape and size of the room on sounds, the properties of materials covering walls and ceilings, and so on. Here too the perception of sounds by ear is meant. Musical acoustics investigates musical instruments and conditions for their best sound.

Physical acoustics deals with the study of sonic vibrations proper. In recent decades, acoustic vibrations beyond the audible limits have been studied intensely (ultrasonics). [This branch of acoustics employs various methods of converting mechanical vibrations into electric oscillations and back, viz. the so-called electroacoustic methods.]

Problems in physical acoustics dealing with sonic vibrations include the analysis of physical phenomena responsible for various properties of sound that can be distinguished by ear.

For example, we distinguish between musical sounds (singing, whistling, sounds of strings and tinkling) and noise (various kinds of cracks, knocks, thunder, hissing and creak). It would not be quite correct to say that musical sounds are produced by musical instruments and noises are not. There exist percussion musical instruments (like drums, kettledrum, castanets, and so on). On the other hand, we speak of whistling bullets, howling wind, humming high-tension cables, broning aeroplanes, etc. recognizing in these sounds certain musical features. Then what is the difference between vibrations producing the effect of a musical sound and noise vibrations?

Before answering this and some other questions of this type, let us consider musical sounds in greater detail. We start with these sounds since they are simpler than noises. This follows even from the fact that a combination of a large number of musical sounds produces the effect of a noise, but a musical sound can never be obtained as a combination of noises.

### 2.3. Musical Tone. Loudness and Pitch

Using a mirror scanning, it was verified that vibrations of a tuning fork are close in form to a harmonic oscillation (see Secs. 1.4 and 1.5). A departure from periodicity due to damping is small for tuning forks, i.e. the amplitude decreases slowly during a long interval of time.

The sound heard when the source emitting it performs a harmonic oscillation is termed a *musical tone*, or just *tone*.

Therefore, the sound of a tuning fork gives a good idea about a tone.

In each tone, we can distinguish by ear two properties: its *loudness* and *pitch*.

Simple observations show that the *loudness of a tone of a given pitch is determined by the amplitude of vibrations*. The sound of a tuning fork gradually attenuates in a certain time after the stroke. This occurs due to damping of vibrations, i.e. a decrease in the amplitude. If we strike the tuning fork stronger, i.e. impart a larger amplitude of vibrations, we shall hear a louder sound than after a weak stroke. This is observed for a string and any other source of sound.

Which property of vibrations determines the pitch of a tone?

If we take several tuning forks of different size, we can easily arrange

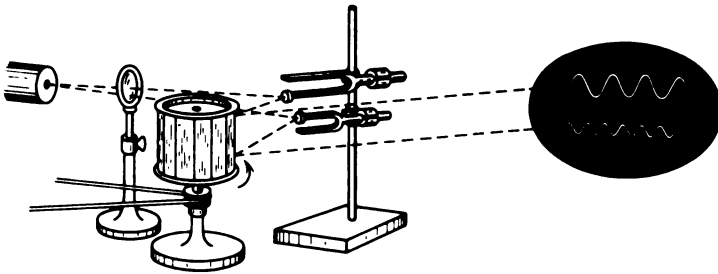


Fig. 37.

Comparing the frequencies of tuning forks.

them by ear in the order of rising pitch of sound. Thus they will turn out to be arranged in the order of decreasing size: the largest tuning fork produces the lowest sound, while the smallest tuning fork, the highest sound. Using a mirror scan, we can easily demonstrate that the smaller the tuning fork, the higher the frequency of its vibrations. Figure 37 shows the schematic diagram of such an experiment.

Thus, the *pitch is determined by vibrational frequency*. The higher the frequency, and hence the shorter the period of vibrations, the higher the sound heard by us.

#### 2.4. Timbre

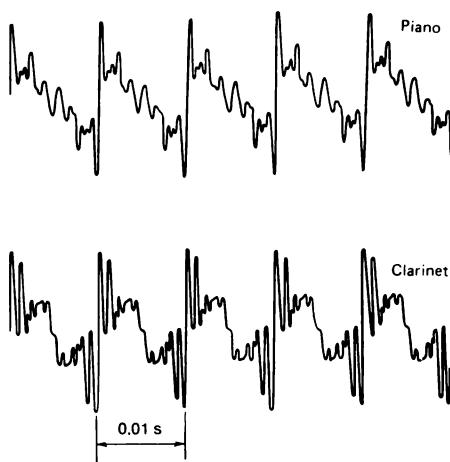
The same conclusions as those drawn at the end of the previous section can be arrived at if instead of tuning forks we use a simple siren, viz. a rotating disc with holes through which an air jet is blown (Fig. 35). Increasing the pressure of the jet, we enhance the oscillations of air density behind the holes. The sound becomes louder, retaining its pitch. Increasing the speed of rotation of the disc, we reduce the period of interruption of the air jet. The sound becomes higher, its loudness remaining unchanged. We can make two or more rows of holes in the disc with different numbers of holes in each row. The sound produced by blowing air jet through each row is the higher the larger the number of holes in a row, i.e. the shorter the period of interruption.

However, when we take a siren as a source of sound, we obtain a periodic but anharmonic vibration: the air density in intermittent jet changes abruptly in pulses. Accordingly, the sound of the siren does not resemble the sound of a tuning fork, although it is a musical sound. We can choose the pitch of the siren sound to make it the same as the one produced by the tuning forks, i.e. to make the siren sound in *unison* with the tuning fork. We can also make the loudness of the two sounds the same. But nevertheless, the sound on the siren can easily be distinguished from the sound of the tuning fork.

Thus, if a vibration is anharmonic, it has one more property in addition to the loudness and pitch. [This is a specific colour of the sound, or its *tembre*.] Due to different *tembres*, we can easily distinguish the sounds of voice, whistling, the sound of a piano string, a violine string, flute, accordion, etc. even if all these sounds have the same pitch and loudness. We distinguish the voices of different persons by their *tembre*.

Which property of vibrations determines the *tembre* of a sound?

Figure 38 presents the oscillograms of acoustic vibrations produced by a piano and a clarinet at the same note, i.e. the sound of the same pitch corresponding to a period of 0.01 s. [The oscillograms show that the period of both oscillations is the same, but they differ considerably in the form, and consequently (see Sec. 1.17) have different harmonic spectra.] The two sounds consist of the same harmonic oscillations (tones) but these tones (fundamental and overtones) are represented in the sounds by different amplitudes and phases.]



**Fig. 38.**

Oscillograms of the sounds of a piano and a clarinet.

Thus, we must find out what parameters are responsible for a certain *tembre*: the amplitudes of harmonics, their phases or both.

An analysis of this problem revealed that only the frequencies and amplitudes of the tones constituting a sound are essential for the human ear, i.e. the *tembre of the sound is determined by its harmonic spectrum*. [The shifts of individual tones in time, i.e. phase shifts of the tones are not perceived by the ear although they may considerably change the form of the resultant vibration.] Thus, the same sound can be heard with quite different forms of vibrations. It is only important that the spectrum, i.e. the frequencies and amplitudes of component tones, remain unchanged.]

Figure 39 represents the spectra of the sounds whose oscillograms are shown in Fig. 38. Since the pitches of the sounds are the same, the frequencies of the tones (fundamental frequency and higher harmonics) are the same. However, the amplitudes of individual harmonics in the two spectra are different. In the sound produced by the piano, harmonics up to the 18th are noticeable, while the 15th and 16th harmonics are practically absent. In the sound produced by the clarinet, harmonics up to the 12th are observed in the spectrum, while the second and fourth harmonics are absent.

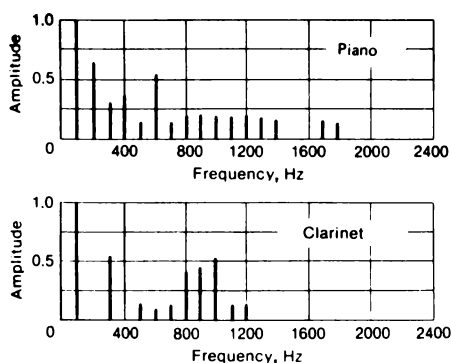


Fig. 39.

Acoustic spectra of the piano and clarinet.

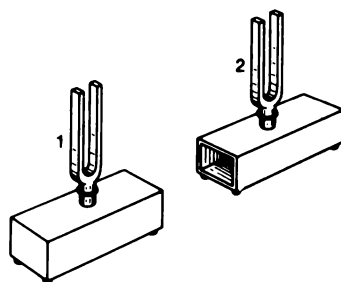


Fig. 40.

Resonance of two tuning forks.

## 2.5. Acoustic Resonance

Resonance phenomena can be observed for mechanical vibrations of any frequency, and in particular, for acoustic vibrations. An example of sonic, or acoustic resonance is provided by the following experiment.

We put one by one two tuning forks so that the open faces of the boxes on which they are fixed are against each other (Fig. 40). These boxes are required to amplify the sound of the tuning forks. The amplification is due to the resonance between a tuning fork and the air column in the box. For this reason, the boxes are called resonators. The functions of these boxes will be described later when we shall analyze the propagation of sonic waves in air. In the experiment considered below, the role of the boxes is purely auxiliary.

Let us strike one of the tuning forks and then clutch it by the hand to quench the vibrations. We shall hear the sound of the second tuning fork.

We can now take two different tuning forks, i.e. having different pitches, and repeat the experiment. In this case, each tuning fork will not respond to the sound produced by the other tuning fork.

The results of the experiment can easily be explained. Vibrations of tuning fork 1 exert through air a certain force on the tuning fork 2, making

it perform forced vibrations. Since the tuning fork 1 vibrates harmonically, the force exerted on the tuning fork 2 varies at the vibrational frequency of the tuning fork 1. If the frequency of the force is the same as the natural frequency of the tuning fork 2, the resonance is observed, i.e. the tuning fork 2 starts vibrating vigorously. If the frequency of the force differs from the natural frequency of the tuning fork 2, its forced vibrations will be so weak that they are inaudible.

[Since tuning forks have a very low damping, their resonance is sharp (see Sec. 1.14).] Therefore, even an insignificant difference in the frequencies of the tuning fork does not allow a tuning fork to respond to vibrations of the other tuning fork. It is sufficient, for example, to attach a piece of plasticine or wax to the prongs of one of two identical tuning forks, and the tuning forks will be detuned, i.e. no resonance will be observed.

Therefore, all phenomena involved in forced vibrations of tuning forks proceed in the same way as in the experiments with forced oscillations of a load on a spring (see Sec. 1.12).

[If a sound is a *note* (periodic vibration) but is not a *tone* (harmonic vibration), this means that it consists of the sum of tones: the lowest tone (fundamental frequency) and higher harmonics.] A tuning fork will be in resonance every time when the frequency of the tuning fork coincides with a frequency of one of the harmonics constituting the sound. Such an experiment can be made with a simple siren and a tuning fork, if we arrange the opening of the resonator of the tuning fork against the intermittent air jet. If the frequency of the tuning fork is 300 Hz, it is easily seen to respond to the sound of the siren not only at 300 interruptions per second (this will be the resonance with the fundamental frequency of the siren), but also at 150 and 100 interruptions, which corresponds to the resonance with the second and third harmonics, and so on.

An experiment similar to that with the set of pendulums (see Sec. 1.16) can easily be carried out with acoustic vibrations. For this purpose, we must have a set of acoustic resonators, viz. tuning forks, strings of organ pipes. Obviously, the strings of a piano form just such a set containing a large variety of oscillatory systems with different natural frequencies. [If we open a piano, press on the peddle and sing a note loudly above the strings, we shall hear the instrument to respond with a sound of the same pitch and a similar *tembre*.] In this experiment, the human voice exerts through air a periodic force on all the strings. However, only those strings respond which are in resonance with the harmonic vibrations (fundamental frequency and higher harmonics) which constitute the note sung by us.

Thus, the experiment with acoustic resonance may also serve as a brilliant illustration of the validity of the Fourier theorem.

## 2.6. Recording and Reproduction of Sounds

We are used to the fact that a human voice or music is reproduced in a telephone, record player, tape recorder or radio set, i.e. a metal plate (or membrane) replaces a rich human voice-producing system and even a whole choir or orchestra. But actually we deal here with acoustic imitation. How can it be obtained?

A modern record player is a modified phonograph created more than 100 years ago by the American inventor Thomas A. Edison (1847-1931). The phonograph was quite a simple device. Air vibrations caused by a source of sound made vibrate a membrane and a needle fixed to it. The needle carved a groove of variable depth on a wax-coated cylinder. Figure 41 shows a magnified cross section of the needle in a groove. Actually, the profile of the bottom of the groove is the oscillogram of vibrations of the point of the needle. If we now place a needle at the initial point of the groove and rotate the wax-coated cylinder again, the uneven bottom of the groove will cause the same vibrations of the needle and the membrane as those during the recording of the sound (the same membrane was used in a phonograph for recording and reproduction of sounds). The vibrations from the membrane will propagate through the air, and reproduced sounds will be heard.

Later, the recording of sound on the basis of transverse oscillations, i.e. in the form of a groove with transverse strokes, replaced the recording with a groove of varying depth. On modern records, the sonic groove has the shape of a spiral along which a needle moves, as the record is rotated on a player, from the edge of a record to the centre. Kinks of this groove can easily be seen through a magnifying glass (Fig. 42).

Leaving aside technical modifications (the choice of materials for records, their manufacturing methods, recording technique, and so on), we emphasize the main problem facing any kind of instruments reproducing sound.

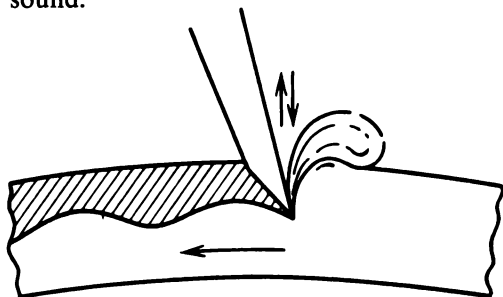


Fig. 41.

Sound groove carved by a needle of a phonograph.

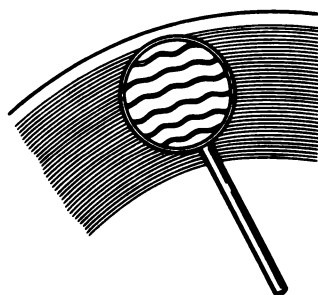


Fig. 42.

Sound track on a modern record.



In recording and reproducing of sounds, we encounter a number of reversible transformations. The recording on a record consists in transforming air vibrations into the vibrations of a membrane and a needle, while the needle leaves traces on the record plate. The recording of sound on cinema film or magnetic tape (in a tape recorder) involves a still larger number of transformations.

The problem is to introduce as little distortions as possible into the spectrum of vibrations, i.e. to preserve the *tembre* of the primary sound. Strong distortions in the vibrational spectrum change the sounds of musical instruments or voice (sometimes beyond recognition) can make the speech *incomprehensible*, and so on.

For perception of sounds, it is immaterial what makes air vibrate (either a membrane or a few tens of instruments in a large orchestra). It is only important that in both cases the vibrations having the same spectrum reach one's ear.

## 2.7. Analysis and Synthesis of Sound

Using a set of acoustic resonators, we can determine the tones constituting a given sound and their amplitudes in the given sound. [The determination of the harmonic spectrum of a complex sound is known as its *harmonic analysis*.] Formerly, such an analysis was actually carried out with the help of a set of resonators, for one, Helmholtz' resonators. These instruments are hollow spheres of various size with a spout that can be inserted into the ear and an opening on the opposite side (Fig. 43). Like the performance of the resonator box of a tuning fork, the operation of such a resonator will be explained below (see Sec. 5.9). For an analysis of sound, it is essential that every time the sound under investigation contains the tone having the frequency of a resonator, the latter amplifies this sound considerably.

However, these methods of analysis are inaccurate and cumbersome. At the present time, they are replaced by more perfect, accurate and high-speed electroacoustic methods. In these methods, an acoustic vibration is first transformed into an electric oscillation having the same form and hence the same spectrum (see Sec. 1.17). Then the electric oscillation is analyzed with the help of electrical methods.

Let us point out one more important result of the harmonic analysis concerning the sound of a human voice. The voice of a person can be distinguished by its *tembre*. But what will be the difference in acoustic vibrations produced by a person who sings the vowels *a*, *e*, *o*, *u* at the same note? In other words, what will be the difference in the periodic vibrations of the air caused by the human voice at different positions of lips and tongue and varying shape of the cavities of the mouth and throat? Obviously, there should be some peculiarities in the spectra of vowels, typical of each vowel,

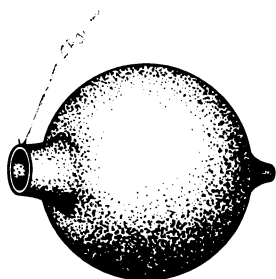


Fig. 43.

Helmholtz' resonator.

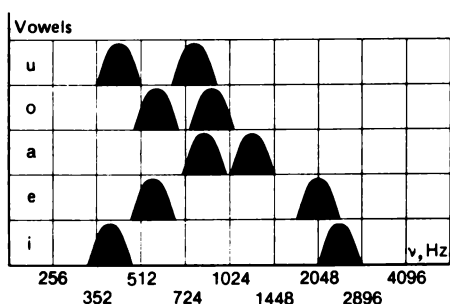


Fig. 44.

Arrangement of formants of vowels on the frequency scale.

besides the peculiarities determining the timbre of the voice of a given person. The harmonic analysis of the vowels confirms this assumption. Namely, the vowels are characterized by the presence of the regions of higher harmonics having larger amplitudes, these regions lying for the same vowel at the same frequencies irrespective of the pitch of the vowel produced by a person. These regions of strong higher harmonics are known as *formants*. Each vowel has two formants typical of it. Figure 44 shows the arrangement of formants for the vowels *u*, *o*, *a*, *e*, *i*.

If the spectrum of a sound in particular, the spectrum of a vowel is reproduced artificially, a person will have the impression of this sound even in the absence of its "natural source". Such a synthesis of sounds (and synthesis of vowels) can be carried out most successfully with the help of electroacoustic devices. Electrical musical instruments make it possible to alter easily the spectrum of a sound, i.e. to vary its timbre. A simple manipulation makes a sound to resemble the tune of a flute, or violine, or human voice and imparts to it peculiar features which distinguish the sound from those of ordinary musical instruments.

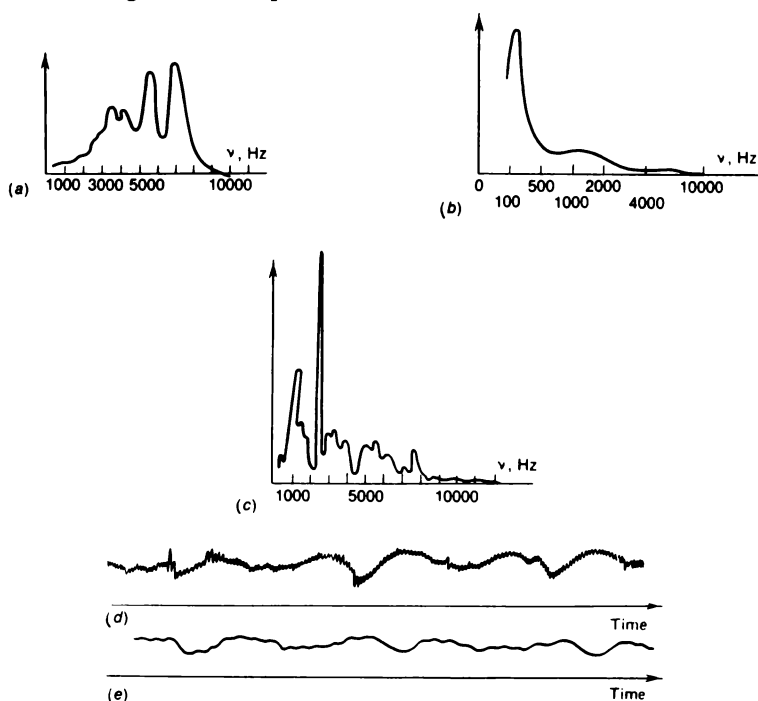
## 2.8. Noises

When a record is played many times, the needle of a player destroys the sonic groove so that its edges are crushed and the shape is distorted. As a result, the worn-out record hisses. A crack in a record produces crackling sounds during playing. Thus, all types of irregularities of the sonic groove, which cause small or large leaps of the needle, produce a noise. This example allows us to outline the main features which distinguish musical vibrations from noise vibrations although there is no sharp border between them.

We hear a musical sound (note) when the vibration is periodic. For example, a piano string produces a sound of this type. If we simultaneously strike several keys, i.e. make several notes sound, the perception of a musical sound will be preserved, but the difference in *consonating* and *dissonating*

notes will become obvious. A *consonance* is a combinations of notes which produces a sound, pleasant to a human ear while a *dissonance* causes an unpleasant sensation. It turns out that the notes whose periods are to one another as small integers are in a consonance. For example, if the periods are as 2 : 3 (such a combination of sounds is known as *quinta* in the theory of music), 3 : 4 (*quarta*), 4 : 5 (*major third*), and so on, a consonance is observed. On the contrary, if the periods are to each other as large integers (say, 19 : 23), we have a dissonance, viz. a musical but unpleasant sound. At a still larger deviation from periodic vibrations, which can be observed when many keys are struck simultaneously with the help of a ruler placed on a key board, the obtained sound resembles a noise.

Noises are characterized by a strongly aperiodic form of vibrations. This can be either a long-term and very irregular vibration of a complex form (hissing or squick), or individual spikes (crackling or knock). From this point of view, the noises should include the sounds corresponding to consonants (hissing sounds, labial consonants). Figure 45 represents the examples of oscillograms and spectra of noise vibrations.



**Fig. 45.**

Acoustic spectra of the consonant "s" (a), Bunsen burner (b), vacuum cleaner (c), and oscillograms of exhaust noise of a motor without a silencer (d) and with a silencer (e).

In all cases, *noise vibrations consist of a very large number of harmonic oscillations at different frequencies*. We obtain such sounds by increasing the number of piano keys struck simultaneously.

Thus, the spectrum of a harmonic oscillation consists of a single frequency. The spectrum of a periodic oscillation contains a set of frequencies, viz. the fundamental frequency and higher harmonics multiple to it. In consonating combinations of sounds, we have a spectrum containing a few such sets of frequencies, such that the fundamental frequencies are to one another as small integers. For a dissonating combination of sounds, the fundamental frequencies are not related through such simple ratios. The larger the number of frequencies in a given spectrum, the closer we approach a noise. Typical noises have spectra containing an extremely large number of frequencies.

## Chapter 3

# Electric Oscillations

### 3.1. Electric Oscillations and Methods of Their Observation

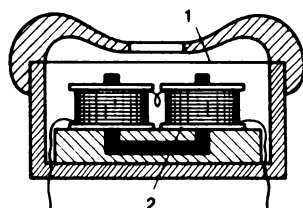
So far, we considered mechanical vibrations which constitute a form of motion of bodies. The problems involving this type of oscillations can be solved in the framework of kinematics and dynamics. However, it was mentioned above (see Sec. 1.2) that along with mechanical vibrations and oscillatory systems, there exist electric oscillations and oscillatory systems. It can be said that they are even more important for engineering than mechanical vibrations. This chapter is devoted to electric oscillations.

The question arises: what oscillates in this case and in which way?

Oscillations can be performed by electric charges on the capacitor plates, electric current in conductors, electromotive force at generator terminals, voltage across a resistance, and so on. In other words, electrical quantities oscillate. By “oscillations” we mean that these quantities do not remain constant but vary with time. But just as not every mechanical motion is a vibration, so not every time variation of electrical quantities is an electric oscillation.

It was shown above that for mechanical vibrations the repetitive nature of a motion, viz. its periodicity is essential. The same property characterizes electric oscillations also. If a physical quantity, say, current, varies periodically, repeating itself, we shall call this variation an *electric oscillation*. An example of such a process is an alternating current in the lighting system, which varies according to a harmonic law (see Vol. 2, Secs. 17.2 and 17.3).

We cannot perceive directly electric oscillations in the same way as we can see the oscillations of a pendulum or hear the vibrations of a tuning fork. But as was mentioned above, both electrically charged bodies and current-carrying conductors interact with one another with certain forces. The measurement of these forces forms the basis for the measurement of electrical quantities: charges, currents, voltages, etc. (see Vol. 2, Secs. 2.14 and 14.4). The mechanical motion of electric motors is caused by these forces (see Vol. 2, Secs. 18.5, 18.6, and 18.7). Using the same (electrostatic or electrodynamic) forces, it is possible to convert, by some method or other, electric oscillations into mechanical vibrations.



**Fig. 46.**  
Telephone (schematic diagram): 1 — membrane and  
2 — electromagnet.

One of such methods employs the attractive force of an electromagnet and is used, for example, in the telephone of an electromagnetic loudspeaker. Figure 46 shows the schematic diagram of a telephone. A current is passed through the winding on an electromagnet whose poles are arranged under the central part of a membrane, viz. a circular iron plate fixed at the edge. Oscillations of the current cause the oscillation of the attractive force exerted on the membrane. As a result, the membrane performs forced vibrations.

If the core of the electromagnet does not possess a permanent magnetization, i.e. attracts the membrane only when a current flows in the winding, the telephone will considerably distort the sound. As a matter of fact, the membrane will be attracted to the core for any direction of the current in the winding, and hence the period of the force acting on the membrane is equal to half the period of the alternating current in the winding. In order to avoid this, electromagnets with permanently magnetized core are used. In this case, the force of attraction of the membrane for a certain direction of the current in the winding will be stronger than in the absence of the current, and for an opposite current it will be weaker. Thus, the period of attractive force now will be the same as the period of the current. Naturally, in these conditions too the conversion of electric oscillations into mechanical vibrations is not free of distortions: the form of vibrations of the membrane repeats the form of oscillations of the current not exactly. However, the possibility of practical employment of such electroacoustic devices (like telephone or loud speaker) just follows from the fact that distortions can be made sufficiently small.

Connecting a telephone or a loud speaker to a lighting system (via the resistance of 100-200 k $\Omega$  since the voltage of 220 V is too high for these instruments), we shall hear a noise, viz. the "sound" of the city current. The vibrations of the membrane due to oscillations of this current have the same frequency (50 Hz), and hence are acoustic vibrations.

Another method of converting current oscillations into mechanical vibrations is based on the rotation of a current-carrying coil in a magnetic field. This principle is used in the design of the loop (mirror-galvanometer) oscillograph (see Vol. 2, Sec. 17.2).

A narrow light loop (mirror galvanometer) is able to follow very fast oscillations of the current, up to 20 kHz, but for higher frequencies an oscillograph with a smaller inertia is required. Such an instrument is a cathode-ray oscillograph. In this device, oscillations are reproduced by the

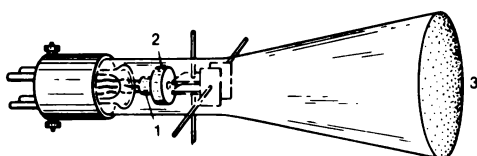


Fig. 47.

Cathode-ray oscillograph. For the sake of clarity, the leads from control plates are passed through the tube walls. Actually, they are connected to the stems of the base (on the left).

motion of a fast electron beam. The schematic diagram of a cathode ray oscillograph is shown in Fig. 47.

Electrons are emitted by a filament cathode *1* and are accelerated on their way to anode *2* due to a voltage (of a few hundred or thousand volts) applied between the cathode (–) and anode (+). The electrons fly in the form of a narrow beam through an opening in the anode (the glass tube is naturally evacuated to a high vacuum). A screen is coated with a substance which glows (fluoresces) under the impacts of the electrons. Thus, the electron beam creates on the fluorescing screen *3* a bright spot. The electric charge transferred by the electron beam to the screen gradually leaks through the inner surface of the glass tube back to the cathode. In order to facilitate and accelerate this leakage of charge, the inner tube walls are coated with a layer of conducting material (graphite).

Behind the anode, there are two pairs of metal *control* plates, the vertical and horizontal. If a constant voltage is applied to any of these pairs, the electric field between the plates will deflect the electron beam either in the vertical or in horizontal direction.

If an alternating voltage is applied to the first (horizontal) pair of plates, the bright spot on the screen will oscillate up and down, reproducing the periodic variation of the applied voltage. Suppose that uniformly increasing voltage is now applied to the other pair of plates. This voltage makes the spot move over the screen in the horizontal direction, say, from left to right. To make the spot that has reached the right part of the screen return very quickly to the left part and be able to repeat its motion, the voltage should be abruptly reduced to the initial value and then uniformly increased again. Such a saw-tooth voltage (Fig. 48) ensures the periodic passage of the spot over the screen, i.e. plays the same role as the rotating mirror in a mechanical oscillograph. For this reason, it is termed *sweep voltage* or *sweep*. As a result, the electron beam draws on the screen the oscillograph of the oscillation applied to the first pair of plates.

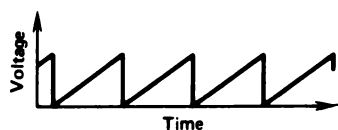


Fig. 48.

Saw-tooth voltage used for the sweep of oscillations.

The inertia of electrons is extremely small, and hence the electron beam is able to follow very fast oscillations up to  $10^9$  Hz. The limit is set by the time of flight of the electrons through the control plates: the electrons successfully follow the variations of the voltage if the voltage across these plates cannot change considerably during the time of their flight.

### 3.2. Oscillatory Circuit

We have an alternating current in the lighting system since generators at electric power plants produce an alternating electromotive force. It was shown earlier (see Vol. 2, Sec. 18.1) that such an electromotive force is induced in a wire frame uniformly rotated in a magnetic field, the period of this emf being determined by the angular velocity of rotation of the frame.

Thus, oscillations of current in a circuit are due to oscillations of electromotive force acting in the circuit in the same way as forced vibrations of a body are caused by vibrations of external mechanical force. Current oscillations are hence forced oscillations.

However, there exist electric circuits in which free electric oscillations may take place, i.e. oscillations in the absence of an external periodic emf. In other words, there exist electric oscillatory system. We shall consider now a simple electric system of this type, viz. an *oscillatory circuit*. This is the term applied to the circuit obtained by connecting a capacitor to an inductance coil (Fig. 49a).

The electrical properties of this circuit are determined by the capacitance  $C$  of the capacitor, the inductance  $L$  of the coil and the resistance  $R$  of the circuit (which is mainly the resistance of the coil). In equilibrium, there is no current in the circuit, and the capacitor is not charged. In order to initiate free oscillations, we must disturb the equilibrium state in some way or other, viz. charge the capacitor or induce a current and then leave the circuit to itself. In Fig. 49a, the circuit is brought out of equilibrium by imparting an initial charge to the capacitor. For this purpose, a battery and a switch are used.

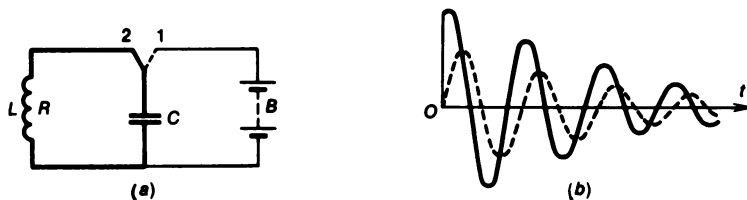


Fig. 49.

(a) Circuit diagram of oscillatory circuit, (b) oscillograms of voltage across the capacitor (solid line) and of current (dashed line) in the circuit.



In the first position of the switch (position 1 in Fig. 49a), the circuit is opened and the capacitor is connected to the battery which charges it to the same voltage as that across the battery terminals. Moving the switch to position 2, we disconnect the battery and close the circuit. From this moment onwards, free electric oscillations take place in the circuit: the charge (and voltage) on the capacitor alternately change sign, passing through their zero values as shown in Fig. 49b by the solid curve. The current in the circuit (dashed curve in the figure) varies similarly but with a phase shift: the current assumes the zero value approximately at the moments when the voltage across the capacitor assumes the maximum positive and negative values.

The smaller the resistance  $R$  of the circuit, the smaller the damping of oscillations and the more exact the coincidence of the passage of the current through zero and the moments of the maximum values of the voltage across the capacitor. In the ideal case when the resistance is absent completely, the oscillations of the current and voltage would be represented by two sinusoids shifted by a quarter of a period relative to each other. In order to clarify some basic regularities, we more than once in the above analysis turned to the ideal oscillatory system in which energy losses are absent. Let us now consider electric oscillations in an ideal oscillatory circuit, i.e. the one with zero resistance. (It should be recalled that in this case undamped oscillations are referred to as natural oscillations.)

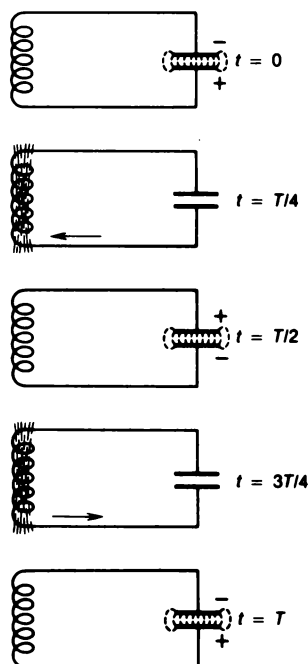
How and why do these oscillations of current and voltage occur? These questions can easily be answered if we recall that the magnetic field cannot vanish or emerge immediately. Indeed, any change in the magnetic field is accompanied by the emergence of an induced emf causing an induced current in the wires. In accordance with Lenz's law, the direction of this current is such that the magnetic field produced by it tends to compensate for the change in the magnetic field responsible for the electromagnetic induction (see Vol. II, Sec. 15.2). This induced magnetic field slows down the variation of the original field, preventing from its instantaneous vanishing or emergence. Thus, magnetic field has an inertia similar to the inertia of a body. A body cannot immediately come to a halt or be shifted since this would signify the existence of an infinitely large acceleration, and hence, according to Newton's second law, would require an infinitely large force.

When we connect a charged capacitor to a coil, at the initial instant the voltage across the capacitor has the maximum value, while the current in the circuit is equal to zero. However, from this moment charges start to move, going over from one capacitor plate to the other, and hence the current causing electromagnetic induction appears. According to what has been said above, the magnetic field, and hence the current producing it,

cannot immediately assume their maximum values but will increase gradually. Since the current carries charges from one plate to the other, the voltage across the capacitor gradually decreases (the capacitor is being discharged). Thus, an increase in the magnetic induction is accompanied by a decrease in the electric field strength. This is what should be expected since, according to the energy conservation law, an increase in the magnetic field energy must be accompanied by a decrease in the energy of the electric field. Therefore, when the voltage across the capacitor vanishes and electric energy turns zero, the magnetic energy attains its maximum value. At this moment, both the current and the magnetic induction in the coil will have the maximum values.

Since the magnetic field (and hence the current) cannot vanish at once, the charge will continue to be transferred in the same direction, and the capacitor will be recharged so that the plate which was initially negative will acquire a positive charge, and vice versa. The current will attenuate and at a certain moment vanish. At this moment, the capacitor will be again charged to the maximum voltage, but now with the opposite sign.

Then the current will flow in the opposite direction so that ultimately the capacitor will be recharged again, i.e. we shall return to the initial state which we had at the moment of closing the switch. Figure 50 shows five



**Fig. 50.**

Oscillations in an oscillatory circuit (see Fig. 49). The state of the circuit is shown every quarter of period  $T$  from the moment of connection of charged capacitor.

consecutive states of an oscillatory circuit separated in time by a quarter of a period, the last diagram (corresponding to a complete period) being the same as the first one. Dashed lines indicate electric field lines in the capacitor and the magnetic field lines in the coil.

### 3.3. Mechanical Analogy. Thomson Formula

If we compare Fig. 50 with Fig. 17 which represents oscillations of a body suspended on springs, we can easily note a similarity in all stages of the process. We can compile a sort of a "glossary" with which the description of electric oscillations can immediately be interpreted in terms of mechanical vibrations, and vice versa. This glossary is as follows.

| <i>Mechanical vibration</i>                            | <i>Electric oscillation</i>                    |
|--------------------------------------------------------|------------------------------------------------|
| (1) mass of a body                                     | (1) inductance of a coil                       |
| (2) elasticity of a spring                             | (2) capacitance of a capacitor                 |
| (3) displacement of the body from equilibrium position | (3) charge on the capacitor                    |
| (4) velocity of the body                               | (4) current                                    |
| (5) potential energy                                   | (5) electric energy<br>(electric field energy) |
| (6) kinetic energy                                     | (6) magnetic energy<br>(magnetic field energy) |

Let us now read again the previous section with the help of this "glossary". At the initial moment, the capacitor is charged (the body is deflected), i.e. an electric (potential) energy is imparted to the system. Then a current starts to flow (the body acquires a velocity) and in a quarter of a period the current and the magnetic energy attain their maximum values, while the capacitor is discharged, and the charge is equal to zero (the velocity of the body and its kinetic energy are maximum, the body passing through the equilibrium position), and so on.

It should be noted that the initial charge of the capacitor, and hence the voltage across it, are produced by the electromotive force of the battery. On the other hand, the initial deviation of the body is created by an external force. Thus, the force acting on a mechanical oscillatory system plays the same role as the electromotive force acting on an electric oscillatory system. Therefore, our glossary can be supplemented by one more entry:

|           |                         |
|-----------|-------------------------|
| (7) force | (7) electromotive force |
|-----------|-------------------------|

The similarity of the two processes can be extended still further. Mechanical vibrations damp due to friction: in each oscillation, a part of energy is converted into heat due to friction, and hence the amplitude becomes smaller and smaller. Similarly, during each reaching of the capacitor,

a part of the energy of the current is converted into heat liberated due to the resistance of the coil. Therefore, electric oscillations in the circuit also attenuate. *The resistance plays the same role for electric oscillations as friction for mechanical vibrations.*

In 1853, the English physicist W. Thomson (Lord Kelvin, 1824-1907) proved theoretically that natural electric oscillations in an oscillatory circuit consisting of a capacitor  $C$  and an inductance coil  $L$  are harmonic oscillations with a period expressed by the formula

$$T = 2\pi\sqrt{LC}$$

( $L$  is expressed in henries,  $C$  in farads, and  $T$  in seconds). This simple and very important expression is known as the *Thomson formula*. Oscillatory circuits including a capacitance and an inductance are often referred to as the *Thomson circuits* since Thomson was the first to describe theoretically oscillations in such circuits. Recently, the term “ $LC$ -circuit” has been introduced (as well as  $RC$ -circuit,  $LR$ -circuit, and so on).

Comparing the Thomson formula with the formula determining the period of harmonic oscillations of an elastic vibrator (see Sec. 1.9),  $T = 2\pi\sqrt{m/k}$ , we note that the mass  $m$  of the body plays the same role as the inductance  $L$  of the coil, while the rigidity  $k$  of the spring plays the role of the quantity reciprocal to the capacitance ( $1/C$ ). Accordingly, in our “glossary” we can write the second entry as follows:

(2) rigidity of the spring      (2) the quantity reciprocal to the capacitance of the capacitor

By choosing different  $L$  and  $C$ , we can obtain any period of electric oscillations. Naturally, depending on the period of electric oscillations, we must employ different methods for their observation and recording (oscillographing). If, for example, we take  $L = 0.5$  H and  $C = 0.5 \cdot \mu\text{F}$ , the period is

$$T = 2\pi\sqrt{0.5 \times 0.0000005} = 0.0031\text{s},$$

i.e. the oscillations occur at a frequency of about 320 Hz. This is an example of electric oscillations with a frequency in the acoustic range. These oscillations can be heard with the help of a telephone and recorded with the help of a loop oscillograph. A cathode-ray oscillograph makes it possible to obtain a sweep of such oscillations as well as of oscillations with a higher frequency. In radio engineering, high-frequency oscillations at a frequency of the order of  $10^6$  Hz are often used. A cathode-ray oscillograph allows us to analyze the form of such oscillations not worse than it was done in the analysis of oscillations of a simple pendulum with the help of the trace left by the pendulum on a darkened plate (see Sec. 1.3).

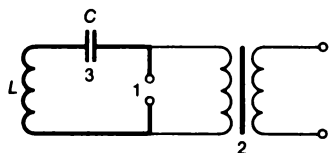
Oscillography of free electric oscillations is not normally used for a single excitation of an oscillatory circuit. As a matter of fact, the equilibrium

state in the circuit sets in after a few periods, or at best in a few tens of periods (depending on the relation between the inductance  $L$ , capacitance  $C$  and resistance  $R$  of the oscillatory circuit). For instance, if in the above example with the oscillatory circuit having a period of 0.0031 s the process of damping is completed in 20 periods, the entire flash of free oscillations will last only for 0.06 s and a visual observation of the oscillogram would be rather difficult in this case. The problem is solved easily when the entire process (from the excitation of oscillations to their complete damping is periodically repeated. Applying the sweep voltage with a period synchronized with the process of excitation of oscillations, we make the electron beam "draw" the same oscillogram in the same region on the screen many times. With a sufficiently high repetition frequency, the pattern observed on the oscillograph screen will be perceived as nonintermittent, i.e. we shall observe a stationary and permanent curve like the one shown in Fig. 49*b*.

In the circuit with the switch shown in Fig. 49*a*, the process can be multiply repeated by switching the key from one position to the other.

In radio engineering, there exist much more perfect and fast methods of such switching, based on the circuit with electronic tubes. However, even before the electronic tubes were invented, an ingenious technique for periodically repeating excitations in an oscillatory circuit has been worked out. This method is based on the spark discharge. We shall consider this simple and visual method in greater detail.

An oscillatory circuit is disconnected by a small gap (spark gap 1) whose terminals are connected to the secondary winding of a step-up transformer 2 (Fig. 51). The current from the transformer charges capacitor 3 until the voltage across the spark gap becomes equal to the breakdown voltage (see Vol. 2, Sec. 8.3). At this moment, a spark discharge occurs in the spark gap, which closes the circuit since the column of a highly ionized gas in the spark channel is as good a conductor as a metal. In such a closed circuit, electric oscillations are excited in the same way as described above. As long as the spark gap conducts the current well, the secondary winding of the transformer is virtually short-circuited by the spark so that the entire voltage of the transformer drops on its secondary winding whose resistance is much higher than the resistance of the spark. Consequently, for a well-conducting spark gap, the transformer practically does not supply any energy to the circuit. Since the oscillatory circuit has a certain resistance, a part of electromagnetic energy is spent for the Joule heat and for the processes occurring in the spark, oscillations attenuate, and after a short interval of time the amplitudes of current and voltage become so small that the spark is extinguished. Then electric oscillations terminate. From this moment, the transformer charges the capacitor again until a new breakdown occurs, and the entire process is repeated (Fig. 52). Thus, spark ignition and extinguishing play the role of an automatic switch ensuring the repetition of oscillatory process.



**Fig. 51.**

Circuit diagram for spark excitation of oscillations in an oscillatory circuit.

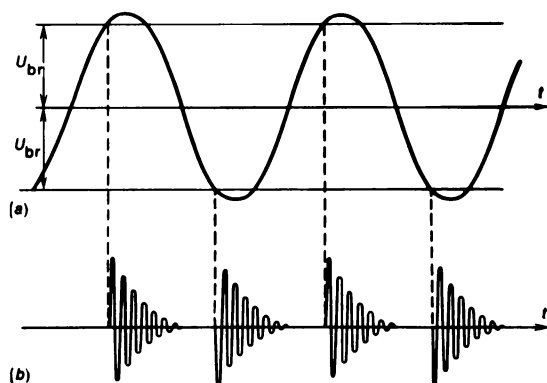


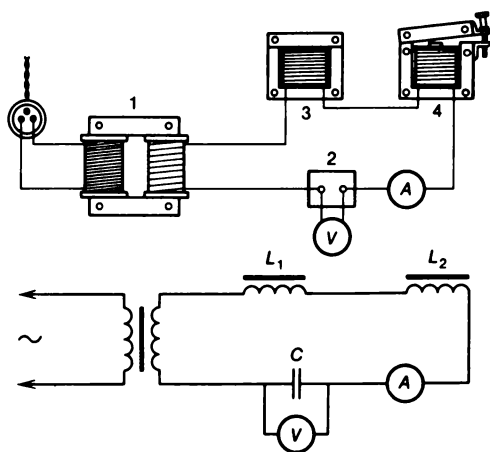
Fig. 52.

Curve (a) shows the variation of the high voltage across the disconnected terminal of the secondary winding of the transformer. At the moments when this voltage attains the breakdown values ( $\pm U_{br}$ ), a spark jumps in the spark gap, the  $LC$ -circuit is closed, and a burst-out of damped oscillations is observed (curves (b)).

### 3.4. Electric Resonance

In the previous section, it was verified that the laws governing free mechanical vibrations and electric oscillations are the same. But in the case of forced oscillations caused by an external periodic force, the similarity is also complete. In the case of electric oscillations, the role of external force is played, as was shown in the previous section, by the electromotive force (abbreviated as emf). Read again Sec. 1.12 where forced vibrations were described, Sec. 1.13 dealing with resonance and Sec. 1.14 devoted to the effect of damping on resonance phenomena in oscillatory systems. What was said there about forced mechanical vibrations completely pertains to electric oscillations. Here too the frequency of forced oscillations in an oscillatory circuit is equal to the frequency of the emf acting in this circuit. The amplitude of forced oscillations is the larger the closer the frequency of the emf to the natural frequency of oscillations in the circuit. When these frequencies coincide, the amplitude attains the maximum value, and electric resonance is observed: the current in the oscillatory circuit and the voltage across its capacitor may strongly exceed their values for detuning, i.e. away from resonance. The resonance phenomena are the more pronounced, the smaller the resistance of the circuit, which hence plays the same role as friction in a mechanical oscillatory system.

All these phenomena can easily be observed by using the alternating current from the lighting system to obtain a harmonic emf and by constructing an oscillatory circuit whose natural frequency can be varied in both directions from the frequency of the current (50 Hz). To avoid very

**Fig. 53.**

A circuit for observing resonance at a frequency of the lighting circuit: 1 — a step-down transformer reducing voltage, say, from 220 to 6 V, 2 — capacitor of capacitance  $C = 1.2 \mu\text{F}$ , 3 — a choke having an inductance  $L_1 = 7.5 \text{ H}$  and a resistance of the winding of  $80 \Omega$ , 4 — a choke with variable air gap having an inductance  $L_2 = 8.3 \text{ H}$  for an air gap width of 2-3 mm, which varies by 15—20% with the gap width on either side of the indicated (resonance) value.

high resonance voltages in the circuit, which may reach a few kilovolts (at a voltage in the mains of 220 V), a step-down transformer should be used.

Figure 53 shows the arrangement of devices and electric circuit diagram for the experiment (notation is the same in the schematic diagram and in the circuit diagram). The circuit includes a step-down transformer 1, capacitor 2 and chokes 3 and 4 (inductance coils with iron cores which ensure the required high inductance). For the convenience of tuning, its inductance is the sum of the inductances of two individual coils. The tuning is carried out by varying smoothly the width of the air gap in one of the chokes (4) within 2-4 mm, thus varying the total inductance. The wider the gap, the smaller the inductance of the coil. In the caption to Fig. 53, approximate values of these quantities are indicated. The voltage across the capacitor is measured with the help of an a.c. voltmeter  $V$ , while an a.c. ammeter  $A$  makes it possible to observe the current in the circuit.

Experiments reveal the following facts. When the inductance of the circuit is small, the voltage across the capacitor slightly exceeds the emf induced in the circuit, i.e. is of the order of a few volts. Increasing the inductance, we note that the voltage increases. This increase becomes the sharper, the closer we approach the resonance value of the inductance. For the numerical values indicated in Fig. 53, the voltage becomes higher than 60 V. A further increase in the inductance leads to a drop in the voltage

across the capacitor. The current in the circuit varies in proportion to the voltage across the capacitor and may reach 20 mA.

This experiment corresponds to the experiment with the load on the springs considered in Sec. 1.12 as an example of a mechanical oscillatory system. In that experiment, it was convenient to vary the frequency of the external force. In the experiment with the electric oscillatory circuit, we pass through the resonance tuning by varying the natural frequency of the oscillatory system, viz. the circuit. However, this does not change the essence of the resonance phenomenon.

The role of electric resonance in engineering cannot be overestimated. We shall consider only one example. Essentially, the *entire technique of radio reception is based on resonance*. A large number of radio stations emit electromagnetic waves that induce alternating emf's (electric oscillations) in the aerial of a radioreceiver, moreover each station excites oscillations of a definite frequency. If we were unable to single out the oscillations induced by a radio station we are interested in from this extremely complex mixture of oscillations, no radio reception would be possible. It is electric resonance that helps us in this situation.

We connect an oscillatory circuit to an aerial through, say, an inductance coil as shown in Fig. 54.

The capacitance of the capacitor can be smoothly varied, thus changing the natural frequency of the circuit. If we tune the oscillatory circuit to the required frequency, for example,  $\nu_1$ , the emf corresponding to this frequency will cause strong forced oscillations in the circuit, while all other emf's will produce weak oscillations. Consequently, *resonance makes it possible to tune the receiver to the frequency of a chosen radio station according to our wish*.

Of course, resonance may produce an extremely harmful effects in electrical engineering, like in mechanical engineering, in situations when it

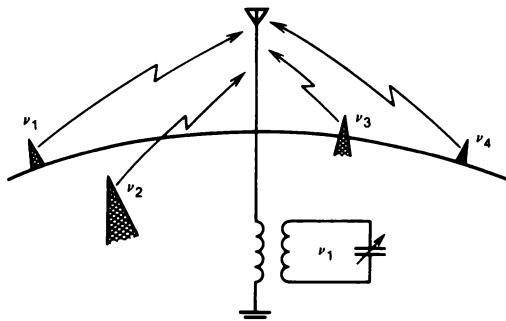


Fig. 54.

Resonance makes it possible to tune to a required station and detune from the other stations. The arrow on the capacitor indicates that its capacitance can be varied.



should be excluded. If an electric circuit is designed for operation in the absence of resonance, the emergence of a resonance may cause a damage: the wires may be heated due to very strong currents, the insulation breakdown may occur due to high resonance voltages, and so on. In the last century, when electric oscillations were studied insufficiently, such damages were frequent. Fortunately, nowadays resonance can be either used or eliminated depending on practical conditions.

### 3.5. Undamped Oscillations. Self-Excited Oscillatory Systems

Free oscillations always attenuate due to energy losses (friction, resistance of the medium, resistance of electric conductor, etc.). However, in physics and in physical experiments there is an urgent need of undamped oscillations which preserve their periodicity all the time the system oscillates. How can such oscillations be obtained? It is well known that forced oscillations in which energy losses are made up for by the work done by an external periodic force are undamped. But how can we obtain an external periodic force? It, in turn, requires a source of certain undamped oscillations.

Undamped oscillations are generated by devices that *may themselves sustain their oscillations* at the expense of a certain permanent source of energy. Such devices are known as *self-excited oscillatory systems*.

Figure 55 shows an example of electromechanical device of this type. A load is suspended on a spring whose lower end is immersed into a cup with mercury during oscillations of this elastic vibrator. One pole of a battery is connected to the upper end of the spring, and the other pole is connected to the cup with mercury. When the load is lowered, the electric circuit is closed, and a current passes through the spring. Due to the magnetic field of this current, the turns of the spring start attracting one

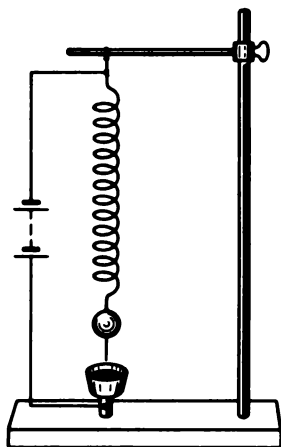


Fig. 55.  
Self-excited vibrations of a load on a spring.

another, the spring contracts, and the load receives an impetus in the upward direction. Then the circuit is disconnected, the turns of the spring no longer attract one another, the load is lowered again, and the entire process is repeated.

Thus, the vibrations of the elastic vibrator, which would be damped had it been left to itself, are sustained by periodic impacts *due to the vibrations of the system itself*. During every impact, the battery supplies a portion of energy which is partially spent for lifting the load. *The system itself controls the force exerted on it* and the energy supply from the source (battery). Oscillations do not attenuate just because the energy given away by the battery during each period is exactly equal to the energy spent during this time due to friction and other factors. As to the period of these undamped oscillations it practically coincides with the period of natural oscillations of the load on the spring, i.e. is determined by the rigidity of the spring and the mass of the load.

Undamped vibrations of the hammer in an electric bell are induced in a similar way, the only difference being that the periodic impacts are produced by a separate electromagnet which attracts an armature fixed to the hammer. Self-excited vibrations at acoustic frequencies can also be obtained. For example, it is possible to induce undamped vibrations of a tuning fork (Fig. 56). When the prongs of a tuning fork move apart, contact 1 is made, a current flows through the winding of electromagnet 2, and the electromagnet attracts the prongs of the tuning fork. Then the contact is broken, and the entire cycle is repeated.

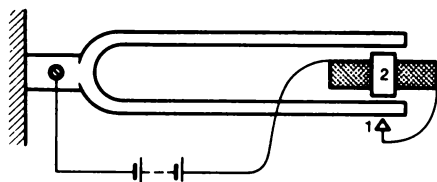


Fig. 56.  
Self-excited vibrations of a tuning fork.

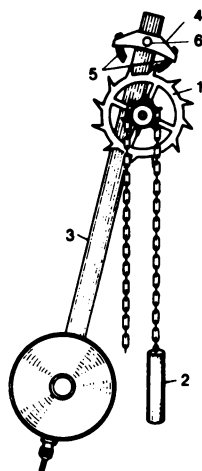


Fig. 57.  
Schematic diagram of a clockwork.

For generating oscillations, the phase difference between an oscillation and the force controlled by it is essential. Let us draw contact 1 from the outer surface of the prong to its inner surface. Now the circuit is closed when the prongs come closer to each other, i.e. the moment when the electromagnet is switched on is shifted by half a period in comparison with the previous experiment. It can easily be seen that in this case the tuning fork will be permanently compressed by the permanently switched on electromagnet, i.e. no oscillations will be observed.

Electromechanical self-excited oscillatory systems are widely used in engineering. Nevertheless, purely mechanical self-excited oscillatory systems are no less used and important. It is sufficient to mention any clock-work mechanism. Undamped oscillations of the pendulum or the balance wheel of a clock are sustained due to the potential energy of the lifted load or elastic energy of a wound spring.

Figure 57 illustrates the operation principle of Galileo-Huygens pendulum clock (see Sec. 1.11). This figure shows the so-called escape wheel. A wheel with bevelled teeth (running wheel) 1 is rigidly fixed to a sprocket through which a chain with load 2 is passed. Pallet 4 is fixed to pendulum 3. Plates 5 on the pallet are bent so that they form segments of a circle with the centre at axle 6 of the pendulum. The pallet does not let the running wheel rotate freely and allows it only to turn by one tooth in every half a period of the pendulum. But the running wheel also acts thereby on the pendulum so that when a tooth of the running wheel is in contact with the left or right plate of the pallet, the pendulum does not receive any impetus and is only slightly decelerated due to friction. But at the moments a tooth of the running wheel "scratches" the endface of the pallet, the pendulum receives an impact in the direction of its motion. Thus, pendulum performs undamped oscillations because it itself allows the running wheel to push it in the required direction. These kicks just make up for the energy losses due to friction. In this case too, the period of oscillations almost coincides with the period of natural oscillations of the pendulum, i.e. is determined by its length.

Self-excited oscillations are observed when a string vibrates under the action of a bow (unlike free oscillations of a string of a piano, harp, guitar and other bowless musical instruments excited by a single kick or jerk). The sounding of wind instruments is also a kind of self-excited oscillations as well as the motion of the piston in a steam engine and many other periodic motions.

A typical feature of self-excited oscillations is that their amplitude is determined by the properties of the system itself and not by the initial deviation or by the impact as in the case of free oscillations. For example, if the pendulum of a clock is deflected too strongly, the energy losses due

to friction will exceed the energy supplied by the running wheel, and the amplitude will decrease. On the contrary, if the amplitude is reduced, the excess of energy supplied to the pendulum by the running wheel will make the amplitude increase. The amplitude corresponding to the balanced energy supply and loss will set up automatically.

### 3.6. Valve Oscillator

In Sec. 8.16 of Vol. 2, while considering electron tubes (valves), it was established that a change in its grid voltage alters the anode current of a triode. When the grid is charged negatively, the electrons cannot fly towards the anode, the current is equal to zero, and the valve is said to be cut-off. Having supplied a positive charge to the grid, we drive the valve to conduction, i.e. a current can pass through it. Variations of the anode current follow the variations of the grid voltage almost instantaneously, the time lag being  $10^{-10}$  s (which is equal to the transit time for electrons flying from the grid to the anode). In other words, an electron valve is a “switch” with a negligible inertia. Therefore, connecting a valve to an oscillatory circuit and a battery so that it is driven to conduction and lets the current flow to the capacitor at the required moments of time, we can obtain an electric self-excited oscillatory system able to excite (generate) undamped electric oscillations.

Obviously, to control the anode current by the oscillations in the circuit, the voltage applied to its grid must depend on the current or voltage in the oscillatory circuit. The oscillatory circuit is then said to be coupled with the grid circuit of the valve. Such an electric coupling can be realized in different ways: with the help of electrostatic induction (capacitive coupling), or with the help of electromagnetic induction (inductive coupling), or by some other method. The most important here is not the way of coupling of the oscillatory circuit with the valve, but the fact that owing to this coupling we observe not only the influence of the valve on the oscillations in the oscillatory circuit, but also the reverse effect of these oscillations of the valve. Various modes of connection of the valve with the oscillatory circuit, which ensure this reverse effect, are examples of the so-called *feedback*.<sup>1</sup> The electric self-excited oscillatory systems of this kind are known as *valve oscillators*. Modern valve oscillators generate oscillations at frequencies to  $10^{10}$  Hz and have diversified applications. They form the basis of any radio station and are included in many types of radio receivers.

---

<sup>1</sup> Generally, a *feedback* is an appliance with which a process occurring in a system (motion or oscillation) is partially used to regulate (control) this process. For example, in steam engines the feedback is executed by a slide-valve mechanism. It starts the engine and controls its operation.

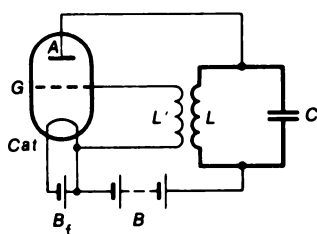


Fig. 58.  
Valve oscillator.

Figure 58 shows one of the numerous circuits of a valve oscillator, viz. the circuit with the inductive feedback.

An oscillatory circuit consisting of an inductance coil  $L$  and capacitor  $C$  is connected in series with a battery  $B$  to the anode circuit of the valve, i.e. between its anode  $A$  and filament (cathode)  $Cat$ . The filament is heated by the current from the filament battery  $B_f$ . The second inductance coil  $L'$  is connected to the grid circuit of the valve between grid  $G$  and cathode  $Cat$ . This coil is inductively coupled with coil  $L$  of the oscillatory circuit. Thus, coils  $L$  and  $L'$  as if form the primary and secondary windings of a transformer without a core. By the way, for oscillators generating low (acoustic) frequencies, a transformer with an iron core can be used.

Coil  $L'$  controls the grid voltage and executes the feedback between oscillations in the oscillatory circuit and on the grid of the valve.

Suppose that oscillations are observed in a circuit consisting of an inductance coil  $L$  and a capacitor  $C$ . An alternating current passing through coil  $L$  induces an alternating emf in coil  $L'$ . The grid alternately acquires a positive and a negative charge relative to cathode  $Cat$ , the period of oscillations of the grid voltage being obviously the same as the period of oscillations in the  $LC$ -circuit, i.e.

$$T = 2\pi \sqrt{LC}.$$

The valve is alternately driven to conduction and cut-off. Therefore, oscillations in the  $LC$ -circuit cause pulsations in the anode current of the valve. The anode current flowing from the anode to the  $LC$ -circuit is branched into the currents through the inductance coil and capacitor (naturally, the constant, i.e. time-independent component of the anode current passes only through the coil since direct current cannot pass through a capacitor, see Vol. 2, Sec. 17.9). If the phase of oscillations of the anode current is chosen appropriately i.e. the "impacts" of the anode current affect the  $LC$ -circuit at the required instants, the oscillations in the  $LC$ -circuit will be sustained (cf. Sec. 3.5). In other words, a portion of energy borrowed from battery  $B$  during each period just makes up for the energy losses in the  $LC$ -circuit over the same period of time, and the oscillations are undamped. If we interchange the terminals of coil  $L'$ , the phase of oscillation of the grid

voltage will be altered by  $180^\circ$ , and no oscillation will be excited (in analogy with the case represented in Fig. 56).

Oscillations can be observed with the help of a cathode ray oscillograph or (if oscillations have an acoustic frequency) with the help of a loudspeaker connected directly to the anode circuit of the valve. We can also connect an incandescent lamp to the capacitor branch of the *LC*-circuit (a flash-light lamp or a motor-car lamp, depending on the power of an oscillator). Since the lamp is connected in series with the capacitor, the direct component of the anode current cannot pass through it. Consequently, the lamp will glow only in the presence of electric oscillations in the *LC*-circuit.

Using a valve oscillator similar to that described above we can easily observe electric resonance if we inductively couple the *LC*-circuit with another oscillatory circuit with a varying capacitor and an incandescent lamp. By varying smoothly the capacitance in this circuit, we can tune it in resonance with the oscillator frequency. With an appropriate choice of the lamp and the coupling between the two *LC*-circuits, we can easily attain conditions in which the lamp glows brightly at resonance and does not glow at detuning.

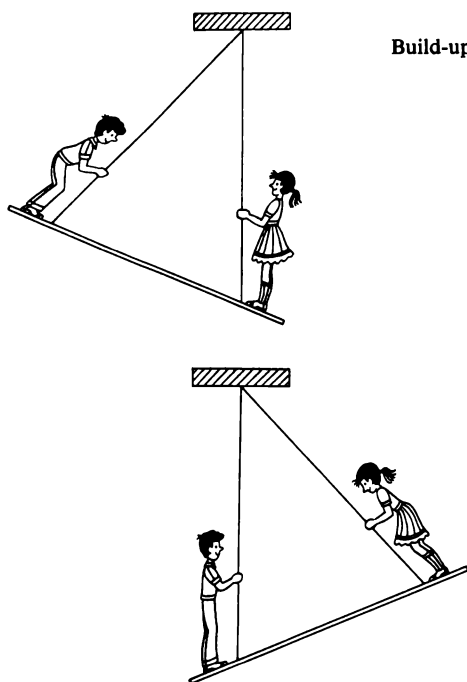
### 3.7. Theory of Oscillations

We began the analysis of oscillations with mechanical vibrations. Then we showed that acoustic phenomena, viz. those perceived by human ear, are essentially based on mechanical vibrations which differ from oscillations of a pendulum only in higher frequencies. After this we considered electric oscillations. In our discussion, we tried to emphasize that the laws governing all these phenomena are strikingly similar.

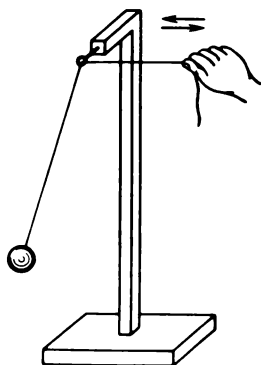
Why did we combine such different phenomena?

All the above chapters were devoted to the theory of oscillations which unifies phenomena not because they are similar in physical nature, but according to universal laws governing them. For example, the regularities observed for free and forced oscillations, resonance phenomena and self-induced oscillations are the same in mechanics and in electricity.

The existence of such similar laws governing quite unlike phenomena from completely different branches of physics is very important for the investigation into natural phenomena. It allows us to use the knowledge gained in the study of phenomena in a certain branch of physics (say, mechanics) for better understanding of phenomena from a completely different branch, say optics. In some cases, this facilitates the investigations, in other cases, this leads to the discovery of new phenomena. The theory of oscillations takes advantage of such a method of investigations whenever it can be applied. Let us consider an example illustrating what has been said about the theory of oscillations.



**Fig. 59.**  
Build-up of swing oscillations.



**Fig. 60.**  
Experiment on parametric  
build-up of oscillations of a  
simple pendulum.

It is well known that oscillations of an ordinary swing can be built up without any impact from outside. For this purpose, the persons standing on the board should alternately squat and rise (Fig. 59). Each partner squats once during a period of oscillations of the swing, and since they do it in turn, it turns out that the centre of mass of the pendulum (swing) is lowered and raised *twice* during a period. This method of exciting oscillations differs in principle from those considered earlier: the *oscillatory system* (swing in our example) *rocks at its natural frequency due to the fact that a quantity determining the period of the system* (in the case under consideration, this is the distance between the point of suspension and the centre of mass) *changes at a double frequency*. Oscillations are excited and maintained at the expense of the work done in changing the period of the system.

This method of exciting oscillations can easily be realized with a simple pendulum, viz. a ball suspended on a string. The string should be passed through a fixed wire ring. Its free end must be periodically pulled by hand, making the load move up and down, i.e. periodically reducing and increasing the length of the pendulum (Fig. 60). If the length of the pendulum varies so that it becomes shorter when the pendulum passes through the vertical position (or near it), i.e. twice during a period, and becomes longer also twice during a period when the pendulum is at extreme positions (or near them), oscillation build-up will be observed, i.e. the amplitude of oscillations will increase. This means that the energy of the oscillating pendulum has increased.

Where does the pendulum get this energy?

In the case under consideration, the energy is received at the expense of the work done by hand muscles. Indeed, making the pendulum shorter by length  $l$  when it passes through the vertical position, we raise the load of mass  $m$  by height  $l$ . If we take into account only the work done against the force of gravity and neglect the work against the centrifugal inertial force, we impart the energy  $mg l$  to the pendulum. The elongation of the pendulum occurs

when it is deflected through the maximum angle  $\alpha$ . In this case, the load will be lowered by the distance  $l \cos \alpha$ , and hence the pendulum will give away the energy  $mgl \cos \alpha$  (Fig. 61). The difference between the received and given away energies equal to  $mgl(1 - \cos \alpha)$  is just the energy imparted to the pendulum each half a period and causing an increase in its amplitude, viz. the build-up. It should be noted that the larger the maximum angle  $\alpha$  (the closer its value to  $\pi/2$ ), the larger the amount of energy received by the pendulum over half a period, i.e. the higher the rate of the build-up.

The same mechanism operates during the build-up of the swing: the energy of the swing increases at the expense of the work done by the swinging partners when they rise (elevate their centre of mass) while passing through the vertical position and squat at the maximum deviation of the swing. Since the effect consists in the change of the length of the pendulum, i.e. the *parameter* determining the period of the system, such an influence is termed *parametric*. It can be seen that the parametric action causes the build-up of oscillations if the frequency of the action is twice as high as the natural (average) frequency of the system.

Let us now go over to another field, viz. the field of electric oscillations. An electric oscillatory circuit obeys the same laws governing oscillations as those for a pendulum. Consequently, if we create in the *LC*-circuit the same conditions as those causing the build-up of oscillations of the swing, electric oscillations should be excited in the *LC*-circuit. Obviously, we must periodically vary a parameter of the circuit which determined its period, i.e. we must vary the capacitance or the inductance, doing it at a frequency twice as high as the natural frequency of the *LC*-circuit. This hypothesis is brilliantly confirmed in experiments. Electric oscillations are indeed excited in this way in the *LC*-circuit.

This method of exciting electric oscillations is used in so-called *parametric oscillators* of alternating current, invented by the Soviet physicists L.I. Mandelshtam and N.D. Papaleksi. Such an oscillator includes an *LC*-circuit consisting of an inductance coil  $L$  and a capacitor whose capacitance  $C$  can be periodically varied during the rotation of its movable part (Fig. 62).

Parametric oscillators can operate with a constant capacitance but with a varying inductance, which turns out to be more convenient from the technical point of view. For high-frequency currents (having a frequency of the order of 100-1000 Hz), they have a number of advantages over ordinary oscillators.

This example clarifies to a certain extent the advantages of the unification of phenomena according to general laws governing them and gives an idea of the scientific value of the theory incorporating these universal laws of oscillations.

Russian and Soviet scientists have made a very large contribution to the theory of oscillations. The works of the outstanding Russian engineer

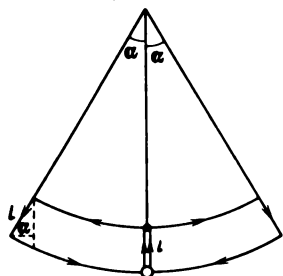


Fig. 61.

To the calculation of work spent on building-up of oscillations over half a period.

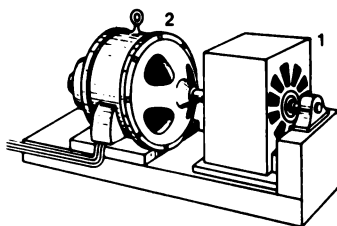


Fig. 62.

Variable capacitor (1) whose movable part is rotated by an electric motor (2). The capacitance of the capacitor varies at a frequency of  $14n$ , where  $n$  is the speed of rotation of the electric motor.



I.A. Vyshnegradsky (1831-1895) on the automatic control of the operation of steam engines, the studies of the founder of the Russian aviation N.E. Zhukovsky (1847-1921) in the theory of aeronautics, the works of the remarkable mathematician A.M. Lyapunov (1857-1918) on the stability of oscillatory motion, the investigations of the founder of seismology B.B. Golitsyn (1862-1916), and the works of the brilliant mathematician and engineer A.N. Krylov (1863-1945) on the theory of rolling and pitching of ships are invaluable not only for special branches to which they directly pertain but also for the universal theory of oscillations. The role of Soviet scientists is not less important since they are the founders of the modern theory of oscillations embracing the theory of self-induced oscillations, parametric excitation of oscillations, the theory of automatic control of operation of machines, and so on. We must mention in this connection the names of the Soviet physicists L.I. Mandelshtam (1879-1944), N.D. Papaleksi (1880-1947), A.A. Andronov (1901-1952) and their pupils, and also our famous mathematicians N.M. Krylov (1879-1955) and N.N. Bogoliubov (b. 1909).

Concluding the chapter, we emphasize once again that acoustic and electromagnetic oscillations, as well as waves which will be studied in the following chapters, naturally differ in their physical nature. The parameters which oscillate are quite different for these fields. It is just the laws governing these oscillatory processes that are similar or universal.

## Chapter 4

# Wave Phenomena

### 4.1. Waves

We shall go over to the analysis of *propagation of oscillations*. In the case of mechanical vibrations, i.e. a vibrational motion of particles of a solid, liquid or gaseous medium, the propagation of vibrations means that vibrations are transferred from some particles of the medium to others. The transfer of oscillations is possible since adjacent regions of the medium are linked with one another. This link can be realized in different ways. In particular, it can be determined by elastic forces emerging due to a deformation of the medium during its vibrations. As a result, a vibration initiated by some way or other in a certain region causes a consecutive emergence of vibrations in other regions which are further and further from the originally excited region, and a so-called *wave* is produced.

Mechanical wave phenomena play a very important role in everyday life. They include the propagation of acoustic vibrations due to elastic properties of the atmospheric air. Owing to elastic waves, we can hear at a distance. The circles spreading over the surface of water on which a stone has been thrown, ripples on the surface of a lake and huge waves in oceans are also mechanical waves of different types. Here the link between neighbouring regions on the surface of water is due to the force of gravity (see Sec. 4.6) or the surface tension (see Vol. 1, Sec. 14.3) rather than due to elastic force. Not only acoustic waves, but also *blast* waves from explosions of missiles and bombs can propagate in air. Seismic stations record vibrations of the ground caused by earthquakes occurring at a distance of thousands kilometers. This is possible since *seismic* waves propagate from the region of an earthquake, which are just vibrations of the Earth's crust.

The waves of quite a different nature, viz. *electromagnetic* waves, are also very important. These waves transport oscillations of electric and magnetic fields produced by charges and currents from some regions of space to others. The link between adjacent regions of an electromagnetic field is due to the fact that any vibration of an electric field induces a magnetic field, and conversely, any change in the magnetic field produces an electric field (Sec. 6.1). A solid, liquid or gaseous medium can strongly affect the propagation of electromagnetic waves, but the presence of such a medium

is not essential for these waves. *Electromagnetic waves can propagate in any space where there exists an electromagnetic field, and hence also in vacuum, i.e. the space free of atoms.*

Light is also one of the phenomena involving electromagnetic waves. Just like a certain frequency range of mechanical vibrations is perceived by the human ear to produce a sensation of a sound, a certain (narrow) frequency range of electromagnetic oscillations is perceived by the human eye to produce a sensation of light.

Observing the propagation of light, we can verify directly that electromagnetic waves can propagate in vacuum. Placing an electric or mechanical bell under the glass bell of an air pump, we see that as the air is being pumped out, the sound gradually attenuates and is finally extinguished. However, the visible pattern of the things under the glass bell and behind it does not undergo any change. This property of electromagnetic waves cannot be overestimated. Mechanical waves do not leave the limits of the Earth's atmosphere, but electromagnetic waves open before us broad avenues of the Universe. Light waves allow us to see the Sun, stars and other celestial bodies separated from us by a huge "empty" space. Using the electromagnetic waves of various frequencies received from these far remote bodies, we can draw very important conclusions concerning the structure of the Universe.

In 1895, the Russian physicist and inventor A.S. Popov (1859-1906) discovered a new and unbounded field of application of electromagnetic waves. He invented a device employing these waves to transmit signals, viz. wireless telegraph. This was the birth of the wireless communication, or radio, due to which a wide range of electromagnetic waves with a much larger wavelength than that of light waves has become of exclusive practical and scientific importance (Sec. 6.7).

The present state of the art of this outstanding invention is such that we have all grounds to speak of radio as one of the miracles of modern engineering. Nowadays, radio makes possible not only the wireless telegraph and telephone communication between any regions of the globe, but also the transmission of visual patterns (television and phototelegraphy), the remote control of machines and missiles (telecontrol), the detection and observation of remote objects that do not emit radio waves (radiolocation). Besides, these branches of engineering can be used to make ships and aeroplanes follow a preset course (radionavigation), to observe the emission of radio waves by celestial bodies, and so on.

We shall consider later some of these applications of electromagnetic waves in greater detail. But even the simple enumeration of these applications, which is far from being complete, confirms their exclusive importance.

In spite of different origins of mechanical and electromagnetic waves, there are many universal regularities typical of any wave phenomena. One of the main laws of this kind is that *every wave propagates from one point in space to another not instantaneously but at a certain velocity*.

#### 4.2. Wave Propagation Velocity

Simple observations convince us that mechanical waves do not propagate instantaneously. Everybody knows that circles on the water surface spread uniformly and gradually as well as sea waves. Here we can observe directly that the propagation of vibrations from one region to another takes a certain time. However, the same can also be observed for acoustic waves that are invisible under ordinary conditions. If a thunderstorm, a shot, an explosion, a whistle of a locomotive, a blow of a hammer, etc. occur at a distance, we first see these phenomena and only after a certain time hear the corresponding sound (Fig. 63). The further is the source of sound from the observer, the longer the time lag. The lapse of time between the lightning and the thunder may sometimes reach a few tens of seconds. If we know the distance from the source of the sound and measure the time lag of the sound, we can determine the velocity of its propagation. This velocity turned out to be 337.5 m/s in dry air at a temperature of 10°C. For the sake of comparison, it should be mentioned that modern aircraft can develop velocities exceeding the velocity of sound in air (so-called ultrasonic velocities), while projectiles shot from a gun fly at a velocity of 1.5 km/s.

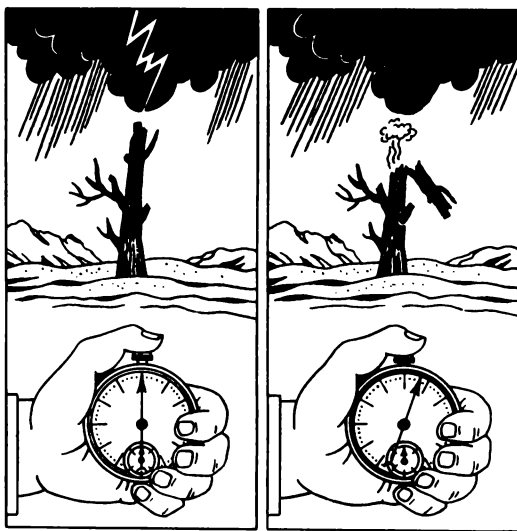


Fig. 63.

We first see a lightning and then hear the thunder.

The velocity of rockets injecting artificial satellites to orbits must be higher than 8 km/s.

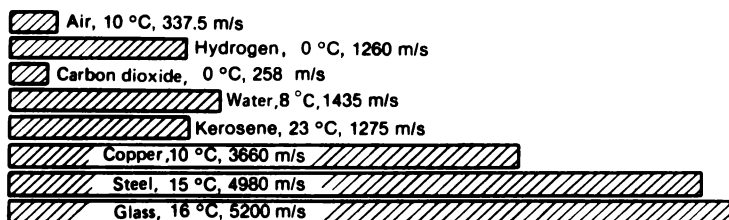
One of the first measurements of the velocity of sound in water was also carried out on the basis of time lag of the sound.

In 1826, Colladon and Sturm made the following experiment on the Geneva lake. Gun powder was ignited on one boat, and at the moment of flash a bell immersed in water was struck with a hammer. The time between the flash and the emergence of sound in a horn also immersed in water was measured from the other boat separated from the first one by a distance of 14 km. The velocity of sound in water at 8°C turned out to be equal to 1435 m/s.

By measuring the time lag of sound in comparison with light, we can obviously obtain the correct value for the speed of sound only if the time of propagation of light can be neglected. Under conditions of conventional observations, this assumption is quite admissible since the measurements show that the *velocity of propagation of light and in general electromagnetic waves in vacuum* (and practically in air) is *approximately equal to 300 000 km/s*.

We see a flash produced at a distance of 3 km with a time lag of just 10  $\mu$ s (a microsecond is one millionth of a second), while the sound covers this distance in about 9 s.

The velocity of acoustic waves is different for different media and, besides, depends on temperature. Modern techniques make it possible to carry out accurate measurements of the velocity of sound with a small amounts of a substance under investigation. Figure 64 shows a diagram representing the velocity of sound in some substances, where the temperature corresponding to a given value is indicated. The figures on the diagram give only an approximate idea about the velocity of sound in a substance since its value also depends on the grade of material (steel or glass) and on the degree of its purity (kerosene).



**Fig. 64.**

Velocity of sound in some gases, liquids and solids.

### 4.3. Radiolocation, Hydroacoustic Detection and Sound Ranging

If we know the velocity of propagation of waves, the measurement of their time lag allows us to solve the inverse problem, viz. to determine the distance covered by the waves.

Negligible intervals of time required for electromagnetic waves to cover distances on terrestrial scale are no longer beyond the limits accessible for observations and can be measured to a high degree of accuracy. This forms the basis of the operation principle of radars, viz. devices for detecting ships, aeroplanes, and other objects.

A radar emits a short electromagnetic signal, viz. a set of very fast oscillations, which lasts for  $1-2 \mu\text{s}$  (Fig. 65). This signal is marked on the screen of a cathode-ray oscillograph as a deflection of the electron beam from the straight line  $AB$  (Fig. 66) along which this beam moves under the action of sweep voltage (see Sec. 3.1). Having been reflected at an obstacle, the signal returns, is received by the radar, amplified and applied to the oscillograph again. The electron beam is again deflected from the straight line  $AB$ , which corresponds to the arrival of the reflected signal. The separation between two spikes on the oscillograph screen on a certain known scale represents the time  $2t$  between the moment of emitting the signal and the moment of arrival of the reflected signal ( $t$  is the transient time in one direction). Since the velocity of propagation of radio waves is known, the straight line  $AB$  can be directly graduated in units of length, and the distance from the reflecting object can be read on the oscillograph screen.

In actual practice, a radar emits not a single signal shown in Fig. 65 but a series of such signals following one another in equal intervals of time many (say, a thousand) times per second. The scan is also made periodic and synchronized with the time of emitting the signals. Thus, the images of the emitted and received (reflected) signals are reproduced on the oscillograph screen many times per second and are perceived by an observer as a permanent pattern.

This is also facilitated by the so-called afterglow of a fluorescent substance used for coating the oscillograph screen. A point on the screen at which the electron beam is incident

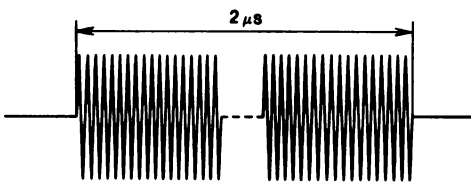


Fig. 65.

A signal (pulse) of a radar shown with a gap since it contains about a hundred of fast oscillations and it would be too extended without a gap.

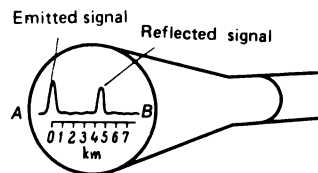


Fig. 66.

Image of signals formed on the screen of the cathode-ray tube of a radar.

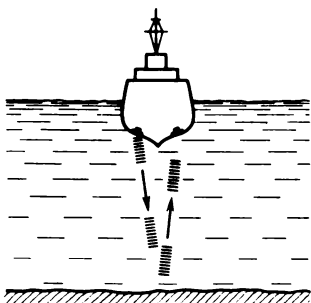
glows for a certain time after the beam is shifted to another region of the screen. The time of afterglow is different for different fluorescent materials. In particular, it can be chosen so that the image formed by the electron beam during a scan period does not disappear before the next period starts, i.e. until the next trace left by the electron beam on the screen.

A periodic repetition of emitted signals produces a permanent and easily observable pattern on the oscillograph screen, which makes it possible to trace the movements of the objects reflecting radar signals. If such an object, say, an aeroplane, moves, the position of the second spike of the electron beam on the screen will change with the distance from the object, i.e. it is possible to determine whether the plane approaches the radar or moves away from it.

Radars can also be used for determining the distance to the shore, and in general to any object capable of reflecting radio waves. Thus, radars can be employed for navigation and for other purposes. At present, radiolocation plays a very important role in various fields, for one, in military trade. The first works on radiolocation in the USSR were started as far back as in 1932 by a group of scientists headed by Yu.A. Korovin. The first radar was constructed in this country in 1939 by Yu.B. Kobzarev and his coworkers.

The problem of measuring distances in water can sometimes be solved by determining the time lag of sound. The time lags observed during propagation of sound are much longer, and hence can be measured more accurately. However, the velocity of propagation of acoustic signal cannot be determined exactly since it is affected by a number of factors: wind, nonuniformity of temperature distribution in a medium (air or water), and so on.

Hydroacoustic detection and echo sounding are based on the same principle (the measurement of the time lag for the reflected signal). Sonars make it possible, for example, to detect submarines from ordinary ships, and vice versa. Using an echo sounder, the depth of the sea bottom can be measured. An echo sounder (sonic depth finder) operates as follows. Special emitter and receiver of ultrasonic waves are mounted at the bottom of a ship (Fig.



**Fig. 67.**  
Operation of echo sounder.

67). These waves are used because they are much shorter than acoustic waves, which ensures certain advantages associated with the directional radiation (see Sec. 4.10). The emitter periodically generates short signals of ultrasonic frequency, while the receiver receives and automatically records on a tape the lagging signals reflected from the sea bottom. In other words, it records the sea depth on a certain scale. As a result, the profile of the sea bottom is recorded during the motion of the ship.

By measuring the difference in the time of arrival of a short sound (like a shot or explosion) at three different sites of observations, one can determine the location of the source of this sound. This method, known as *sound ranging*, was used during the war for notching the location of artillery units of the enemy.

#### 4.4. Transverse Waves in a Cord

Let us now consider mechanical (elastic) waves in greater detail. Their properties are determined by many factors: the link between neighbouring regions of the medium, the size of the medium (for instance, the pattern of propagation of waves in a body of bounded size will be other than in practically boundless medium such as atmospheric air), the shape of the body, and so on.

In this and the following sections we shall consider two types of elastic waves, viz. transverse and longitudinal waves.

Let us suspend a long cord or a rubber pipe at one end. If we rapidly displace the lower end of the cord to one side and return it to the initial position, the bent will “run” upwards along the cord (Fig. 68a), reach the point of suspension, will be reflected and then return to the lower end (Fig. 68b). If we move the lower end of the cord continuously so that it performs a harmonic oscillation, a sinusoidal wave will “run” along the cord (Fig. 68c). It will also be reflected from the point of suspension, but phenomena occurring as a result of this reflection will be considered later (Secs. 5.4 and 5.5).

When we say that a wave or a single bent “runs along the cord”, we briefly describe the following process: each point of the cord performs the same oscillation as the lower end of the cord excited by us, but the oscillations of each point lags behind (in phase) the more, the farther this point from the lower end of the cord. Figure 69 clarifies the kinematics of the process of transfer of oscillations from point to point. Here we see different consecutive phases of this process, beginning from the “equilibrium position” in a quarter of a period. Each numbered circle of the series performs a harmonic oscillation about its own “equilibrium position” with the same amplitude and frequency. The oscillation of each next circle lags behind the oscillation of the previous circle by  $1/12$  of a period (i.e. by  $30^\circ$  in



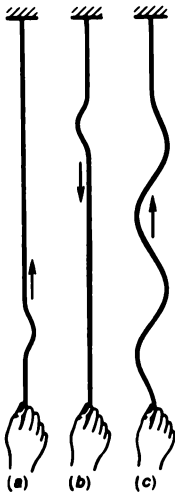


Fig. 68.

Motion of the bent along a cord: (a) the bent runs upwards, (b) returns upon reflection, and (c) sinusoidal wave.

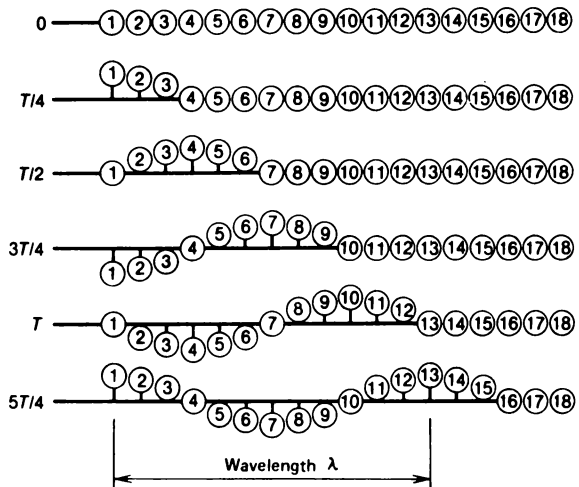


Fig. 69.

Kinematics of a transverse wave.

phase). Thus, circle 4 lags behind circle 1 by  $90^\circ$ , 7 lags behind 1 by  $180^\circ$ , 10 is behind 1 by  $270^\circ$ , 13 lags behind 1 by  $360^\circ$ , i.e. oscillates in the same way as 1. Further, the process is repeated: circle 14 oscillates (when the wave reaches it) in the same way as 2, 15 oscillates as 3, and so on. It can be seen that the wave along which the circles are arranged moves to the right. Here, the wave propagates during a period over a distance equal to the separation between the circles oscillating with the phase shift of  $360^\circ$ , i.e. oscillating similarly (obviously, the phase shift by the number of degrees multiple to  $360^\circ$  is equivalent to the absence of the phase shift).

The distance over which the wave propagates over a period is known as the *wavelength*. Consequently, *the wavelength is the distance between two closest points of a sinusoidal* (or, which is the same, harmonic) *wave oscillating in phase*. The wavelength is usually denoted by the Greek letter  $\lambda$ .

Thus, we have a *dual periodicity in a wave*. On the one hand, each particle of the medium performs a periodic vibration *in time*. On the other hand, all particles lie on a line whose shape is periodically repeated *in space*. The wavelength  $\lambda$  plays the same role relative to the shape of the wave in space as the period  $T$  plays relative to the oscillation in time.

If we are interested in the velocity of propagation  $v$  of the wave, i.e. the distance covered by it per unit time, we must obviously divide the wavelength  $\lambda$  (traversed over the period  $T$ ) by the period  $T$ :

$$v = \frac{\lambda}{T}.$$

If we know two quantities appearing in this formula, the third quantity can be easily calculated.

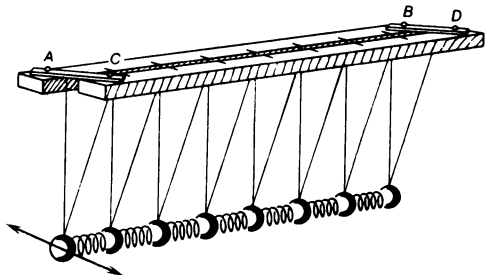
It was mentioned at the beginning of the section and should be emphasized once again that the *propagation of a wave is equivalent to a lagging transfer of oscillatory motion from one point of the medium to another*. There is no mass transfer of a substance of a body in which the wave propagates.

Every point of the cord (like each circle in Fig. 69) vibrates perpendicularly to the direction of propagation of the wave, i.e. across the direction of propagation. For this reason, this type of wave is known as *transverse*.

Why is a vibrational motion transferred from one point of the medium to another and why does this occur with a time lag? To answer this question, we must analyze the dynamics of the wave.

A displacement of the lower end of the cord causes deformation of the cord in this region. The emerging elastic forces tend to eliminate this deformation, i.e. there appear stresses which drag the region of the cord immediately adjoining the region of the cord displaced by hand. The displacement of this second region causes deformation and the state of stress in the next region, and so on. (Of course, in actual practice there are no separate regions of the cord, and the process is continuous.) The regions of the cord have a mass, and due to inertia gain or lose the velocity under the action of elastic forces with a time lag. When the lower end of the cord is deflected by hand to the extreme right position and starts to move to the left, the adjacent region will continue to move to the right, comes to a halt and then also move to the left with a certain time lag. Thus, the *time lag in the transfer of vibrations from one point of the cord to another is due to the elasticity and mass of the material of which the cord is made*.

A simple model can be used to illustrate the effect of these properties. Two laths  $AB$  and  $CD$  (Fig. 70) are hinged to transverse rods  $AC$  and  $BD$ . Balls are suspended from the laths so that each ball hangs on two strings whose upper ends are fixed to  $AB$  and  $CD$  respectively. If parallelogram  $ABCD$  is arranged so that the laths  $AB$  and  $CD$  are in contact (as shown in Fig. 70), the balls will be able to swing only in the planes perpendicular



**Fig. 70.**

A model for demonstrating transverse waves.

to the laths. If we give  $ABCD$  a rectangular shape, the balls will be able to oscillate only in the direction parallel to laths  $AB$  and  $CD$ . (This case is shown in Fig. 74 and will be analyzed only in the next section.) The balls are connected to one another by not very stiff springs.

In this model of an elastic body, viz. a chain of alternating balls and springs, the two properties we are interested in are singled out: the mass is mainly concentrated in the balls and elasticity in the springs. Moving an extreme ball by hand from side to side, we can easily observe the transfer of vibrations from ball to ball by means of the deformations of the springs and the time lag between the vibration of each ball relative to the vibration of the previous ball. As a result, a transverse wave running along the chain is induced (Fig. 71).

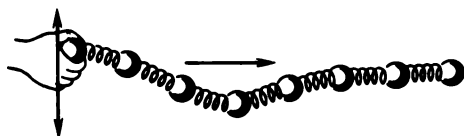


Fig. 71.  
A transverse wave.

The stiffer the springs and the smaller the mass of the balls, the shorter the time lag between vibrations of each ball and those of the previous ball, and hence the larger the wavelength for the same period. But an increase in  $\lambda$  at a constant  $T$  indicates an increase in the velocity of propagation of the wave. Our model hence leads us to the following regularity which is actually observed for elastic bodies: *the velocity of propagation of elastic waves is the higher the larger the stiffness of a body and the lower its density.*

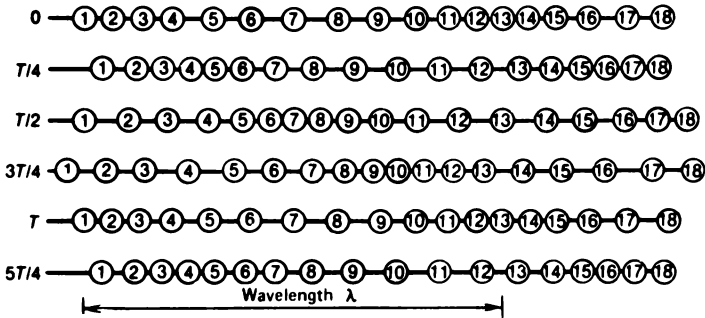
#### 4.5. Longitudinal Waves in an Air Column

Let us now consider another type of waves. We again take a body having an elongated shape, viz. air column contained in a tube in which a piston can move. We make the piston perform harmonic vibrations. What will happen to the air column?

The results obtained in the previous section immediately provide the answer to this question. Here too each region of the body (air layer) has a mass, and a compression of air creates an excess pressure, i.e. the air exhibits elasticity. Consequently, an elastic wave appearing in the air column will run away from the piston (Fig. 72). But now the oscillatory motion in the wave differs from what was observed in the previous case: air particles vibrate in the same direction as the piston, i.e. *along the direction of propagation of the wave*. Such waves are known as *longitudinal*.



Fig. 72.  
A wave in a tube.

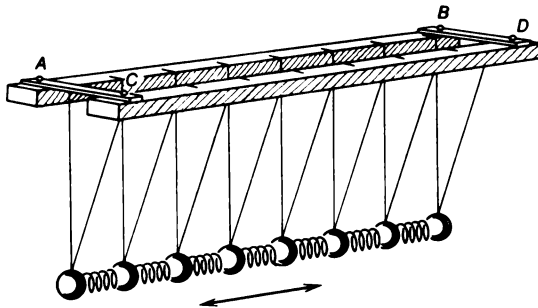


**Fig. 73.**  
Kinematics of longitudinal waves.

Figure 73 explains the kinematics of a longitudinal wave. In this figure, like in Fig. 69, there is a set of numbered circles which vibrate harmonically about their equilibrium positions. As before, the amplitude and frequency is the same for all the circles, but the phase of each circle lags behind the phase of the previous circle by  $30^\circ$ . The difference between this process and the one shown in Fig. 69 is that now the circles vibrate not across the row but *along* it. Besides, Fig. 73 shows a steady-state wave. As a result of these longitudinal oscillations with a time lag from circle to circle, a wave running to the left and consisting of alternating compressions and rarefactions emerges.

The dynamics of a longitudinal wave can be analyzed with the help of the model described in the previous section.

By imparting a rectangular shape to the frame  $ABCD$  (Fig. 74), we allow the balls to swing only in the longitudinal direction, i.e. parallel to laths  $AB$  and  $CD$ . Moving an extreme ball back and forth, we shall see alternating rarefactions and condensations formed and propagating along the chain.



**Fig. 74.**  
A model for demonstrating longitudinal waves.

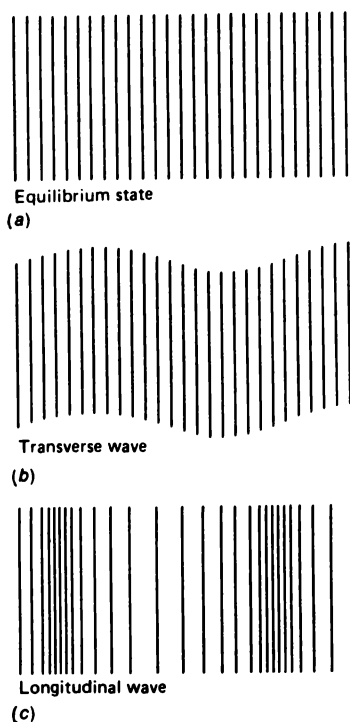


Fig. 75.

Deformation of a medium in a longitudinal and transverse waves.

Just as in this model, longitudinal and transverse waves can propagate in a continuous medium extending in all directions. Transverse waves in such a medium are *shear waves* in which the layers perpendicular to the direction of propagation are displaced during their vibration parallel to one another, i.e. without rarefactions or condensations (Fig. 75b). Longitudinal waves are *compression waves*<sup>1</sup> in which the deformation of layers of a medium consists in a change in their density so that a wave is formed by alternating compressions and rarefactions (Fig. 75c).

Of course, the definition of the wavelength given in the previous section completely remains in force for longitudinal waves also.

For transverse waves, it can be said that the wavelength is equal to the distance between two adjacent crests (or troughs) of the sinusoid, while for longitudinal waves it is equal to the distance between the middles of two adjacent compressions (or rarefactions). The velocity of propagation

of a longitudinal wave is connected with the wavelength and the period through the same relation as for a transverse wave. Of course, this does not mean that the velocity of propagation of the two types of wave in a body is the same. On the contrary, the velocity of compression waves in any medium is higher than that of shear waves (and hence the wavelength of a longitudinal wave is higher than the wavelength of a transverse wave for the same period).

Saying "in any medium", we should stipulate that we refer to any *solid* medium. As a matter of fact, *transverse elastic waves can propagate only in solids, while longitudinal waves can propagate in solids, in liquids and in gases*. Thus, we can compare the velocities of propagation of the two types of waves only in solids, while in liquids and gases only longitudinal waves can be observed.

Why is this so?

<sup>1</sup> A compression can be positive (condensation) and negative (rarefaction).

As was mentioned above, in a *transverse* wave a *shear* of layers relative to one another is observed. But elastic forces emerge during shear only in solids. In liquids and gases, adjacent layers freely slide over one another without the emergence of opposing elastic forces. And since there are no elastic forces, no elastic waves can be formed.

In a *longitudinal wave*, the regions of a body experience *compressions* and *extensions*, i.e. change their volume. Elastic forces emerge during a change in volume in solids as well as in fluids. For this reason, longitudinal waves are observed in bodies in any of the three states of aggregation.

#### 4.6. Waves on the Surface of a Liquid

It was mentioned earlier that the formation of waves can be associated not with the elastic force but with the force of gravity. Therefore, it is not astonishing that the waves propagating over the surface of a liquid are not longitudinal waves. However, they are not transverse either since the motion of particles of a liquid is more complicated in this case.

If the surface of a liquid is lowered at some point (for example, under the action of the force of gravity as a result of a contact with a solid), the liquid will flow into this depression, filling it and forming an annular depression around it. At the outer edge of the annular depression, the particles of the liquid continue to flow down, and the diameter of the ring increases. But at the inner edge of the ring, the particles move up again, forming an annular crest. Behind it, a new trough is formed, and so on. A particle moving down also moves in the backward direction, while rising particles also move forward. Thus, each particle does not simply vibrate in the transverse (vertical) or longitudinal (horizontal) direction, but describes a circle.

In Fig. 76, the dark circles represent the position of the particles on the surface of a liquid at a certain instant, while light circles show the positions of these particles after a certain time, when each particle has traversed a part of its circular trajectory. These trajectories are shown by dashed lines, and the traversed segments, by arrows. The line connecting the dark circles represents the profile of the wave. This figure corresponds to a large amplitude (i.e. the radius of the circular trajectories of particles is not small in comparison with the wavelength), for which the profile of the wave does not resemble a sinusoid: it has wide troughs and narrow crests. The line connecting the light circles has the same shape but is shifted to the right (towards the lag in phase), indicating that the wave has been displaced as a result of motion of particles of the liquid in circular trajectories.

It should be noted that not only the force of gravity, but also surface tension (see Vol. 1, Sec. 14.3) plays an important role in the formation of surface waves. Like the force of gravity, surface tension tends to flatten the surface of the liquid. The propagation of a wave causes the deformation of the surface at each point: convexities are flattened, become concave and then flattened again. As a result, the area of the surface, and hence the energy of surface



**Fig. 76.**  
Motion of particles of a liquid in a wave on its surface.

tension are altered. It can easily be seen that the role of surface tension for a given amplitude is the higher, the larger the curvature of the surface, i.e. the shorter the wave. Therefore, for long waves (low frequencies), the force of gravity plays the major role, but for sufficiently short waves (high frequencies), the surface tension becomes of major importance. Naturally, the boundary between “long” and “short” waves is not sharp and depends on the density of a liquid and the surface tension characterizing it. For water, this boundary corresponds to waves with a wavelength of about 1 cm. This means that for shorter waves (known as capillary waves), the forces of surface tension prevail, while for longer waves, the force of gravity plays a decisive role.

In spite of a complex “longitudinal-transverse” nature of surface waves, they obey the laws governing any wave process and are very convenient for analyzing some of these regularities. For this reason, we shall consider the methods of their obtaining and observation in greater detail.

Experiments with such waves can be made by using a shallow bath with a glass bottom having an area of about  $1 \text{ m}^2$ . A bright electric bulb arranged under the glass at a distance of 1-1.5 m makes it possible to project this “pond” onto a screen or just onto the ceiling (Fig. 77). The magnified shadow image allows us to observe all phenomena occurring on the surface of water. In order to suppress reflections of waves from the edges of the bath, the surface of the edges is corrugated, and the edges are inclined.

We fill the bath with water to a depth of 1 cm and touch its surface with the end of a wire or the sharp end of a pencil. A circular wave will be seen to run from the point of contact in all directions. The velocity of its propagation is not high (10-30 cm/s), and hence its displacement can be easily watched.

If we fix the wire on an elastic plate and make it vibrate so that during each vibration of the plate the end of the wire touches the water, a system of circular crests and troughs will run over the surface (Fig. 78). The distance between two neighbouring crests or troughs, viz. the wavelength  $\lambda$ , is connected with the period  $T$  of vibration of the plate through the well-known formula  $\lambda = vT$ , where  $v$  is the velocity of propagation of the wave.

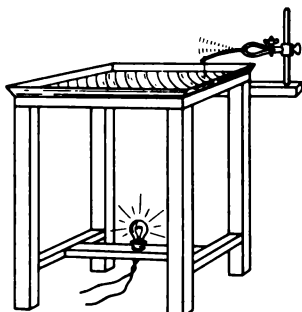


Fig. 77.

A bath for observing waves on water surface.

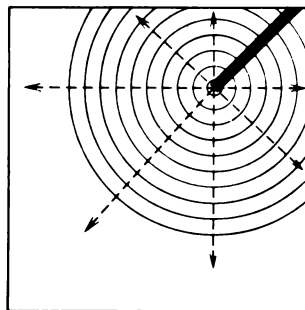


Fig. 78.

Circular waves.

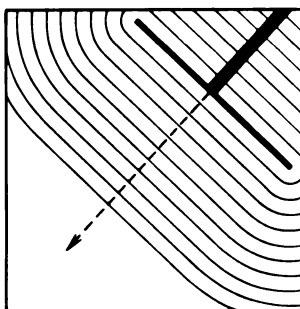


Fig. 79.  
Rectilinear waves.

The lines perpendicular to the crests and troughs indicate the directions of propagation of the wave. For a circular wave, the directions of propagation are obviously represented by straight lines diverging from the centre of the wave as shown in Fig. 78 by dashed arrows.

If we replace the wire by a ruler fixed at its edge so that it is parallel to the surface of water, we can obtain a wave not in the form of concentric rings but in the form of rectilinear crests and troughs parallel to one another (Fig. 79). In this case, the wave propagates from the middle of the ruler in a single direction.

Circular and rectilinear waves on the surface give an idea of spherical and plane waves in space. A small source of sound radiating uniformly in all directions produces in the surrounding space a spherical wave in which compressions and rarefactions of air alternate in the form of concentric spherical layers. A region of a spherical wave whose dimensions are small in comparison with the distance to the source can be treated approximately as a plane wave. Naturally, this refers to waves of any physical origin, both elastic and electromagnetic. For example, any portion (within the surface of the Earth) of light waves received from stars can be considered as a plane wave.

Henceforth, we shall use the above experiment with a water bath more than once since such surface waves are very graphic and convenient for observing the main features of many wave phenomena, including such important effects as diffraction and interference. We shall use the waves in a water bath to get a *general idea* about the processes observed both for elastic (in particular, acoustic) and for electromagnetic waves. Whenever it is possible to observe finer features of waves processes (in particular, in optics), we shall provide a more detailed interpretation for these features at a later stage.

#### 4.7. Energy Transfer by Waves

Propagation of an elastic wave, which involves a consecutive transfer of motion from one region of the medium to another, is hence associated with



an energy transfer. The energy is supplied by the source of the wave when it sets in motion the layer of the medium, which is in direct contact with the source. From this layer, the energy is transferred to the next layer, and so on. Thus, a wave propagating in a medium creates an energy flow diverging from the source. The idea about the energy flow carried by waves was introduced in 1874 by the Russian physicist N.A. Umov (1846-1915). He obtained a formula for calculating the intensity of the wave.

When a wave encounters various bodies, the energy carried by it can be converted into work or transformed to other kinds of energy.

A striking example of such an energy transfer without a mass transfer is provided by blast waves. A blast wave can break window panes and walls, i.e. does a large mechanical work, at a distance of tens of metres from the site of explosion, where neither splinters nor hot air flow can get. However, energy is transferred by very weak waves also. For example, a flying mosquito emits an acoustic wave ("mosquito peep") whose power, i.e. the energy emitted in 1 s, is as low as  $10^{-10}$  W.

If the size of a source is small enough and the energy propagates uniformly in all directions from it, the source can be considered as a *point* source<sup>2</sup>, and the wave diverging from it will be *spherical*. In this case, the energy emitted by the source is distributed uniformly over the spherical surface of the wave. It can easily be seen that the energy per unit surface of this sphere is the smaller, the larger the radius of the sphere. The area of the spherical surface or any region cut on it by a cone (Fig. 80) increases in proportion to the square of the radius. This means that when a distance from the source increases twice, the area increases four-fold, and the wave energy falling per unit area of the spherical surface decreases to 1/4 of the initial value.

The energy transported per second by a wave through a surface element whose area is  $1 \text{ m}^2$ , i.e. the power carried through a unit area, is known

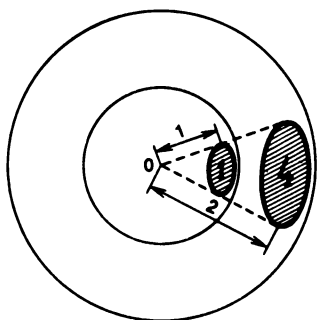


Fig. 80.

As the radius increases twice, the area of the surface increases four-fold.

<sup>2</sup> The conditions under which a source can be treated as a point source will be considered in greater detail when we shall describe light sources (see Sec. 8.2).

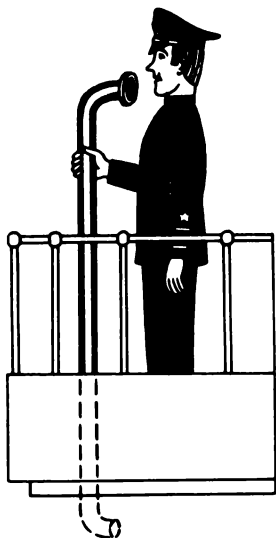
as the *wave intensity*. Thus, the *intensity of a spherical wave decreases in inverse proportion to the squared distance from the source*.

The power of the sound produced by a flying mosquito at a distance of 2 m from it is distributed over a spherical surface whose area is  $4\pi \times 2^2 \approx 50 \text{ m}^2$ , i.e. the intensity of sound at such a distance constitutes about  $10^{-12} \text{ W/m}^2$ . This negligibly small value is close to the audibility threshold and gives an idea about an extremely high sensitivity of the human ear.

If we limit the ability of a wave to diverge in all directions, the decrease in the wave intensity will be smaller. For example, an acoustic wave propagating in a tube does not diverge, and hence preserves a high intensity over a long distance. This principle is used in the construction of speaking tubes which are still in use on small ships for communication between the captain's bridge and the machine department (Fig. 81).

In order to increase the loudness of sounds at large distances, megaphones are sometimes used (Fig. 82). It should be borne in mind, however, that outside the megaphone the divergence of the wave is not restricted any longer, and the reason behind the amplification of the sound is different here: the megaphone concentrates the wave energy within a certain solid angle, i.e. produces a directional radiation (see Sec. 4.10). But within this solid angle, the wave intensity decreases in inverse proportion to the squared distance.

The intensity of a wave propagating in a cylindrical tube should not decrease with increasing distance since the energy is transported here through sections of the same area. However, in actual practice we observe



**Fig. 81.**  
Speaking tube on a ship.



**Fig. 82.**  
Loudspeaker (megaphone) produces directional radiation.

the attenuation due to the energy absorption by the medium in which the wave propagates. At each point of the medium, a part of energy transported by the wave is spent on the work against friction (viscosity) of the medium and is converted into heat. Due to absorption, the intensity of a spherical wave actually decreases at a higher rate than in inverse proportion to the squared distance. During the propagation of the wave in the tube, the energy transported by the wave is also absorbed by the walls of the tube.

Electromagnetic field variations are transferred in the form of electromagnetic waves. Naturally, these waves also transport energy not in the form of the kinetic and potential energy of the particles of the medium but in the form of the energy of the electric and magnetic field. The entire energy received from the Sun and maintaining the life on the Earth has such a form. The total power of electromagnetic waves emitted by the Sun amounts to  $4 \times 10^{23}$  kW. At a distance of 150 million kilometers (which is approximately equal to the separation between the Sun and the Earth), the intensity of electromagnetic waves is equal to  $1.4 \text{ kW/m}^2$ . This value is known as the *solar constant*. Only about 43% of this energy reaches the surface of the Earth because of reflection at clouds, and scattering and absorption by the atmosphere (see Vol. 1, Sec. 18.2).

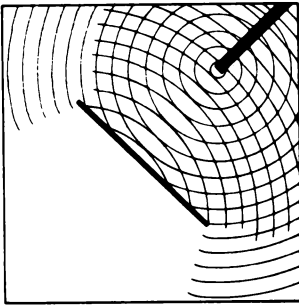
If the Sun had moved away from us to the same distance as that to the nearest star (i.e. to 4 light years), the intensity of its electromagnetic radiation near the surface of the Earth would have been only  $2 \times 10^{-8} \text{ W/m}^2$ . But even if one hundredth of this energy corresponded to the visible light, the intensity of this light would be much higher than the sensitivity threshold of the human eye.

It is interesting to note that the sensitivity threshold of the eye is almost the same as that of the ear. The eye is able to perceive luminous energy fluxes of about  $3 \times 10^{-13} \text{ W/m}^2$ . However, modern radio receivers can compete with the human eye as regards sensitivity: a good radio receiver can “hear” a radio station with the wave intensity of  $10^{-14} \text{ W/m}^2$  at the reception site.

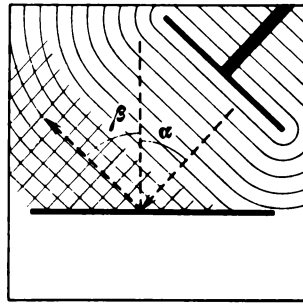
#### 4.8. Reflection of Waves

Let us place a flat plate on the path of the waves in a water bath, the length of the plate being large in comparison with the wavelength  $\lambda$ <sup>3</sup>. We shall see that behind the plate there is a region in which the surface of water is almost at rest (Fig. 83). In other words, the plate forms a *shadow*, viz. the space where the waves do not penetrate. In front of the plate, the waves will be seen *to be reflected* from it, i.e. the waves incident on the plate produce the waves propagating from the plate. These reflected waves

<sup>3</sup> In Figs 83 and 84, as well as in some subsequent figures, the wavelength  $\lambda$  is taken not small enough in comparison with the size of the plate. This is done only for the figures to be distinct.

**Fig. 83.**

A shadow cast by a large plate.

**Fig. 84.**

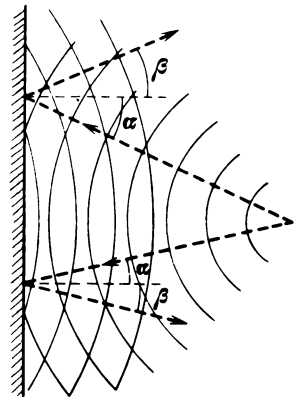
The angle of reflection is equal to the angle of incidence.

have the shape of concentric arcs as if propagating from a centre lying behind the plate. A sort of network is formed in front of the plate by the primary waves incident on the plate and the reflected waves propagating against the incident waves.

How does the direction of propagation of a wave change upon its reflection?

Let us consider the reflection of a plane wave. We denote by  $\alpha$  the angle between the normal to the plane of our "mirror" (plate) and the direction of propagation of the incident wave (Fig. 84), while the angle between the same normal and the direction of propagation of the reflected wave is denoted by  $\beta$ . Experiments show that for any position of the "mirror",  $\beta = \alpha$ , i.e. the *angle of reflection of the wave is equal to the angle of incidence*.

This law of reflection is a universal wave law, i.e. it is valid for any waves, including light and sound waves. The law remains in force also for spherical (or circular) waves, as is shown in Fig. 85. Here the angle of reflection

**Fig. 85.**

The law of reflection is observed at any point of the reflecting plane.

tion  $\beta$  at different points of the reflecting plane is different, but at each point it is equal to the angle of incidence  $\alpha$ .

The reflection of waves at obstacles is a very common phenomenon. The well-known echo is caused by the reflection of acoustic waves at buildings, hills, woods, etc. A multiple echo is heard if acoustic waves reach an observer after consecutive reflections at a number of obstacles. Peals of thunder are also of the same origin. They are multiple repetitions of a strong “crackling” of a huge electric spark (lightning).<sup>4</sup> The methods of detection and ranging described in Sec. 4.3 are based on the reflection of electromagnetic and elastic waves from obstacles. The phenomenon of reflection is observed most frequently for light waves.

A reflected wave is always weaker to a certain extent than the incident wave. A part of energy carried by the incident wave is absorbed by the body at whose surface the wave is reflected. Acoustic waves are well reflected at solid surfaces (plaster or parquet) and much worse by soft surfaces (carpets, curtains, and so on).

Every sound ceases not immediately after its source stops emitting, but attenuates rather gradually. The phenomenon known as *reverberation* (prolongation of sound) is due to the reflection of sound in an empty room. In large empty halls, reverberation is strong, i.e. “booming” or “rumbling” is observed. If, however, there are many soft absorbing surfaces in a room (like easy chairs and sofas, clothes of people, or curtains), no rumbling is observed. In the former case, sound undergoes multiple reflections before the energy of an acoustic wave is absorbed practically completely. In the latter case, absorption is much more rapid.

Reverberation essentially determines the acoustic properties of rooms and plays an important role in architectural acoustics. For a given room (auditorium, concert hall, etc.) and a given type of sound (speech or music), the absorption should be specially chosen. It must be not very strong (otherwise the sound would be muffled and indistinct) and not too weak (otherwise a long reverberation would make the sound of speech or music confused).

#### 4.9. Diffraction

The formation of a shadow is a frequently observed and usual phenomena for light waves. With acoustic waves, the situation is different. It is difficult to create a shield on their path. We hear a sound from behind the corner of a building or while standing behind a fence, tree, and so on. Why do these obstacles not “cast a sonic shadow”?

---

<sup>4</sup> The primary sound caused by a lightning is prolonged. As a matter of fact, a lightning can be extended over a few kilometres, and hence the sound from different regions of the lightning reaches an observer with a time lag.

The following circumstance is important here. The wavelength of an acoustic wave is 33.7 cm for a frequency of 1000 Hz, while it is equal to 3.37 m for a frequency of 100 Hz. Thus, the dimensions of objects around us (with an exception of large buildings) are not large at all in comparison with the wavelength of acoustic waves. At the same time, in order to observe a clear reflection and a formation of shadow in the experiment with the bath described in the previous section, we used an obstacle (plate) of a size much larger than the wavelength  $\lambda$ .

How does the nature of the formed shadow depend on the size of the obstacle?

We place obstacles of different sizes on the path of a linear surface wave in the bath (Fig. 86). It can be seen that when an obstacle is big enough in comparison with the wavelength  $\lambda$  (Fig. 86a), its shadow is comparatively sharp: small disturbances are observed only at the very edge of the shadow, indicating that the wave slightly rounds the edge of the obstacle. As the size of an obstacle is reduced, the shadow becomes less and less sharp (Fig. 86b). When the size of an obstacle becomes comparable with the wavelength, practically no shadow is formed. Figure 86c shows that in this case the wave rounds the obstacle and propagates behind it almost as if there were no obstacle. This *rounding of the edge of an obstacle by a wave*, which is clearly observed when the size of the obstacle is small in comparison with the wavelength, is known as *diffraction*.

The absence of a clear-cut sound shadow under normal conditions is just the result of the diffraction of acoustic waves, which is hence encountered very often. The diffraction of light waves is not so easy to observe as that of acoustic waves since light waves are very short (the wavelengths are of the order of just  $10^{-4}$  mm).

Diffraction is a very important phenomenon that can be observed for any wave process. It will be studied in greater detail in the third part of this volume, devoted to physical optics. It will be shown, in particular, that diffraction does not allow us to distinguish small details of an object observed in an optical instrument (including the human eye). Diffraction

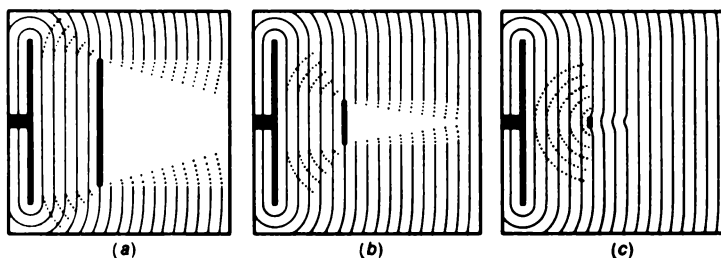
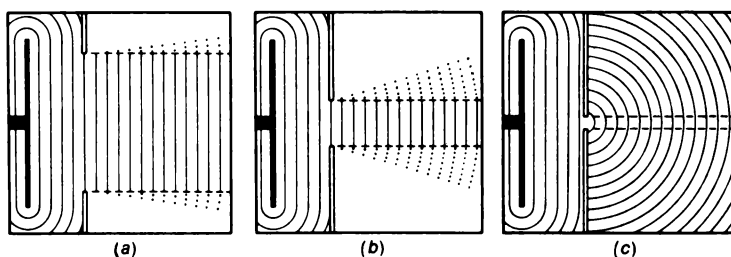


Fig. 86.

No shadow is formed behind a small obstacle.

**Fig. 87.**

Diffraction does not allow to isolate as narrow wave bundle as desired.

makes it impossible to obtain as narrow wave beams as desired with the help of megaphones, mirrors, or orifices in screens (diaphragms). This can be easily demonstrated even for acoustic waves.

Let us arrange two plates on the path of a linear wave in a water bath so that the gap between them isolates a bounded bundle from this wave (Fig. 87a). Bringing the plates closer to each other, we shall see that the bundle isolated by them cannot be made as narrow as desired. As the gap becomes smaller, diffraction (viz. rounding the edges of the plates by the wave) becomes more and more pronounced (Fig. 87b). Finally, when the gap width becomes comparable with the wavelength or even smaller, we obtain behind the plates semicircular waves diverging in all directions from the gap as from the centre (Fig. 87c) instead of a narrow bundle.

#### 4.10. Directional Emission

In experiments with waves in a water bath, we obtained a circular wave with the help of a point touching the surface of water. In order to obtain a linear wave front, the point was replaced by the edge of a ruler. It should be noted that the ruler touching the surface of water must be held so that its edge is parallel to the surface of water, i.e. the vibration should be excited by all points of the edge simultaneously. In other words, in order to obtain a wave with a rectilinear front, a large number of oscillators vibrating in phase must act along the same straight line. If the ruler were arranged at an angle so that some of its regions touched water before others, the nature of the wave would be quite different. Henceforth, we shall also assume that when a wave is excited by an oscillator like a ruler, all points of this oscillator vibrate in phase.

In a circular wave produced by a point, the energy is transferred in all directions, while in a wave with a rectilinear wave front, the energy is transported directionally, viz. in the direction perpendicular to the edge of the ruler. What determines the directionality of radiation?

Let us try to obtain waves with the help of rulers of different lengths acting as oscillators. It can easily be seen that the shorter the edge touching

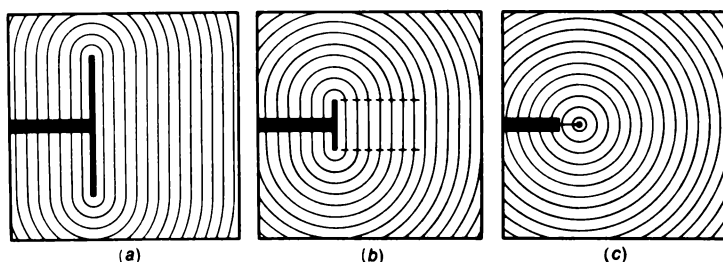


Fig. 88.

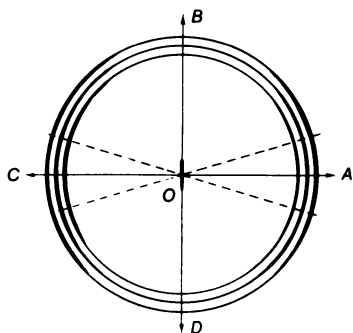
The longer the edge of a ruler, the longer the linear wave front is retained.

the water, the shorter and the less clear is the segment of the rectilinear wave (Fig. 88). This is natural since in emission we in fact also deal with diffraction, but now it is observed around the emitting body itself. In the same way as the nature of diffraction of an incoming wave at an obstacle is determined by the ratio of the size of the obstacle and the wavelength  $\lambda$ , the form of the wave emitted by a ruler depends on the ratio between the length of its edge and the wavelength  $\lambda$ . Comparing the waves obtained from rulers of different lengths with the wave diverging from the gap between two plates, i.e. comparing Figs. 88 and 87, we see that an analogy of the entire pattern is complete and that the effect of the length of the ruler in one case and the width of the gap in the other case on the form of the wave is the same. The larger the ratio of the length  $l$  of the ruler to the wavelength  $\lambda$ , the further from the ruler the linear wave front is retained.

Nevertheless, however long the edge of the ruler may be, we can find a large distance from the ruler, such that the wave becomes circular, its humps and troughs becoming concentric circles.

Does this mean that at such long distances from the oscillator its shape and size do not affect the form of the wave? It turns out that this is not so. The shape of the wave front, and of humps and troughs of the wave does indeed become circular at sufficiently large distances from the wave, but the *intensity* of such a circular wave *will not be the same in all directions*. A completely directionless wave having the same intensity in all directions is formed only when a point or other object small in comparison with the wavelength  $\lambda$  disturbs water surface. If, however, a wave is produced by the edge of a ruler whose length is considerably larger than  $\lambda$ , the wave intensities on continuations  $OB$  and  $OD$  of the edge of the ruler will be lower than that along directions  $OA$  and  $OC$  perpendicular to the edge (Fig. 89) even at large distances where the wave is already circular. The emitted energy is mainly concentrated in a certain sector of a circular wave around directions  $OA$  and  $OC$ , this sector being the narrower (and



**Fig. 89.**

At large distances from the ruler, the wave is circular, but its intensity is not uniform in different directions.

the directionality of the emission the higher), the longer the ruler is in comparison with the wavelength  $\lambda$ . In the case of a point, this "sector" embraces the entire circle, the wave being directionless.

Thus, the larger the length of a linear oscillator in comparison with the wavelength  $\lambda$ , then, firstly, the longer the distance from the oscillator at which the rectilinear wave front is preserved, and secondly, the higher the intensity of the wave concentrated around the direction perpendicular to the oscillator at distances where the wave becomes circular.

These conclusions concerning the waves on the surface of the liquid are applicable also to any waves in space provided that the emitter is modified accordingly. For example, instead of the edge of a ruler, we can imagine a disc (membrane) vibrating in air or under the water surface. All what was said above can be repeated for the longitudinal wave generated by such an oscillator. But now instead of rectilinear and circular waves, we obtain plane and spherical waves respectively. In particular, the concentration of sound with the help of a megaphone, described in Sec. 4.7, can also be explained by the increased size of the oscillator (the opening of the megaphone) in comparison with the size of the mouth, i.e. by the increased ratio of the oscillator size to the wavelength.

## Chapter 5

# Interference of Waves

### 5.1. Superposition of Waves

Let us make the following experiment with the waves in a water bath. We shall make two rulers vibrate on two elastic plates so that by disturbing the surface of the water, they produce two plane waves (Fig. 90). The rulers are arranged at an angle to each other so that wave bundles emitted by them overlap in the region  $aa'b'b$  and then diverge again. Experiments show that the propagation of each wave through region  $aa'b'b$  is completely independent of the presence or absence of the other wave. *One wave does not affect in any way the propagation of the other wave.*

The same refers to acoustic waves: the propagation of a sound from any source does not experience any influence of other acoustic waves propagating in some way or other during the same time through the same region of the medium. The same is true for light waves also: their propagation from any source to the observer's eye and all what is seen by the observer due to these waves do not at all depend on a large number of other light waves crossing the path of light from the object under observation in all possible directions.

What occurs in the regions of space where, like in region  $aa'b'b$  in Fig. 90, waves are superimposed on each other?

Each particle of the medium on the path of a wave vibrates with the period of the wave. If this particle is on the path of two waves, it simultaneously participates in vibrations due to the two waves, i.e. its motion is the sum of these two vibrations. Thus, *the superposition of two (or more) waves is the*

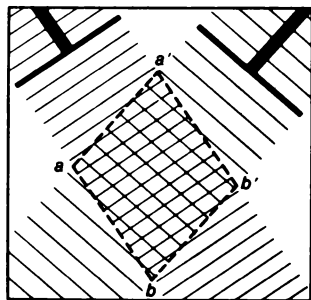


Fig. 90.

Waves generated by two vibrating rulers overlap without affecting the propagation of each wave.

*addition of their vibrations at each point of the medium through which these two (or more) waves propagate.* It was shown above that the superposition of waves in any region does not affect their propagation both in this region and outside it.

## 5.2. Interference of Waves

However, if two waves having the *same frequency*, and hence the *same wavelength*, are superimposed, a peculiar and very important phenomena may occur, which will be discussed now.

We fix two wire points on the same vibrating plate so that they touch the surface of water simultaneously when we push the plate. We shall obtain two circular waves of the same wavelength, which propagate from the two centres and are superimposed in the bath.

If the distance between the points is larger than the wavelength, we shall observe a pattern similar to that shown in Fig. 91. The waves are not just enhanced as it should be expected at first glance, but a more complex phenomenon takes place. There will be regions on the surface of water where vibrations are especially strong (the maxima are:  $aa'$ ,  $bb'$ , ...), alternating with regions where vibrations are suppressed (the minima are:  $mm'$ ,  $nn'$ , ...). Such a pattern of alternating maxima and minima of vibrations is known as *interference pattern*, and the *phenomenon of superposition of waves which results in such a pattern is termed interference*.

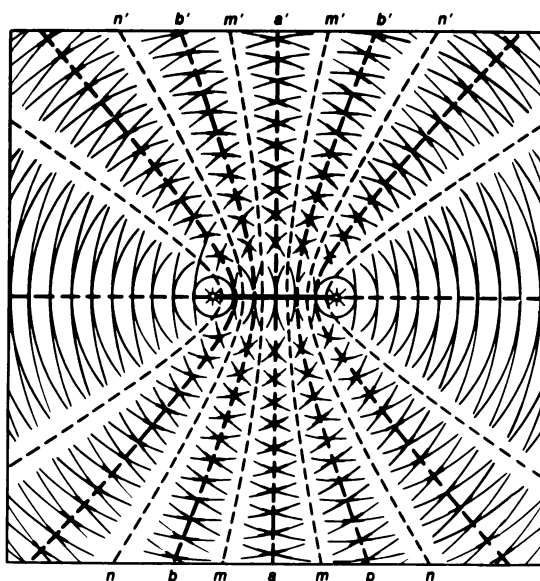


Fig. 91.

The result of superposition of circular waves.

The formation of the interference pattern in this experiment will give us an idea about the conditions for interference in general.

The vibration at each point on the water surface is the sum of vibrations due to each wave separately. Suppose that the crests of the two waves meet at a certain point at a given instant, i.e. the two waves arrive at this point in phase. The level to which water rises at this point is found to be elevated. In half a period ( $T/2$ ), the crests will change for troughs in both waves simultaneously since they have the same period. The surface of water then be lowered. Thus, in this region enhanced vibrations will be observed. On the contrary, at a point where the crest of one wave meets the trough of the other wave, i.e. where the waves are in antiphase, vibrations will suppress each other. Such a suppression will be observed all the time since at any instant the waves will come in opposite phases (in particular, in half a period, the trough of the first wave and the crest of the second wave will be at this point). Therefore, for the emergence of the interference pattern it is essential that the waves propagating from two centres be matched: the phase shift between vibrations in the two waves at a given point remains constant all the time.

If we varied arbitrarily the phase of vibrations of one of the sources, the phases of the two vibrations at any point on the surface of water would then alternately coincide and differ, and the positions of maxima and minima would not be stable. Similarly, if the periods of vibrations of the two waves were different, the enhancement of vibrations at every point of the surface would be changed for the suppression, then the vibrations would be enhanced again, and so on. The larger the difference in the periods or the higher the rate of variation of the phase of one of the vibrations, the more rapidly the maxima and minima change their positions.

When we speak about an interference pattern, we mean a *stable, time-independent pattern* of alternation of maxima and minima. Such a stable pattern emerges only when the waves being superimposed have the same period and constant phase shift of vibrations at each point. Such waves are known as *coherent*.

Consequently, *a stable interference can be observed only provided that the waves are coherent.*

In the case considered above, the coherence is ensured by the coupling of the sources (two wire points) attached to the same vibrating plate.

In such a case, two coherent sources produce waves emerging at the points where they are excited in phase, i.e. crests (or troughs) emerge from the two sources simultaneously. A set-up in which one of the waves lags behind in phase relative to the other wave can be visualized. But if a time lag remains constant during the experiment, the sources are also coherent (although they are not in phase), and the waves generated by them will produce a stable interference pattern. Thus, a constant phase shift of the two

waves is required for their coherence, the absolute value of this shift being immaterial.

### 5.3. Conditions for Formation of Interference

#### Maxima and Minima

Can one predict the location of maxima and minima of vibrations on an interference pattern?

Let us consider Fig. 92 showing the interference pattern for the waves from two coherent sources  $S_1$  and  $S_2$ . Let the two sources vibrate in phase, i.e. the crests (or troughs) emerge simultaneously from them. Obviously, vibration maxima will be observed on the straight line  $aa$  all whose points are equidistant from sources  $S_1$  and  $S_2$ , since the crests (or troughs) of the two waves reach the points lying on this line simultaneously and in phase. Similarly, the vibrations will be enhanced on line  $bb$  whose all points are closer to  $S_2$  than to  $S_1$  by the wavelength  $\lambda$ . At all points of  $bb$ , the wave from source  $S_1$  will lag behind the wave emitted by source  $S_2$  exactly by a period, which means that the phases of the two waves will coincide again. The same will be observed for line  $cc$  whose points are closer to  $S_2$  than to  $S_1$  by  $2\lambda$ , i.e. one wave lags behind the other wave by two periods, as well as for the lines  $b'b'$ ,  $c'c'$ , ..., whose points are closer to  $S_1$  than to  $S_2$  by  $\lambda$ ,  $2\lambda$ , ....

Similar arguments show that the suppression of vibrations (minimum) will be observed on lines  $mm$ ,  $nn$ , ... and  $m'm'$ ,  $n'n'$ , ..., whose all points are closer to one source than to the other by a half-wave ( $\lambda/2$ ), three half-waves ( $3\lambda/2$ ), and in general, by an odd number of half-waves. Indeed, at all points of these lines, a crest of one wave will meet a trough of the other wave, or in other words, the waves come to these points in antiphase.

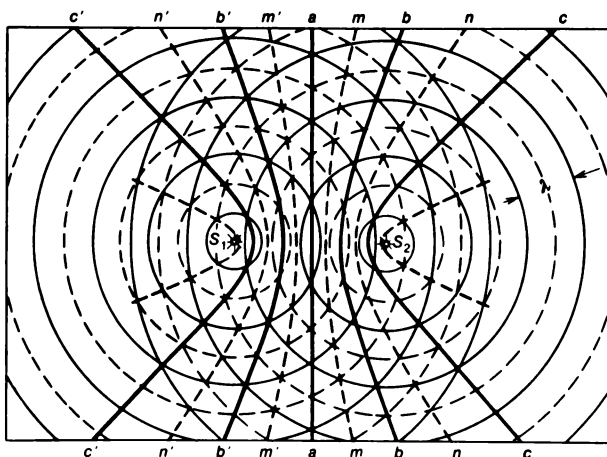


Fig. 92.

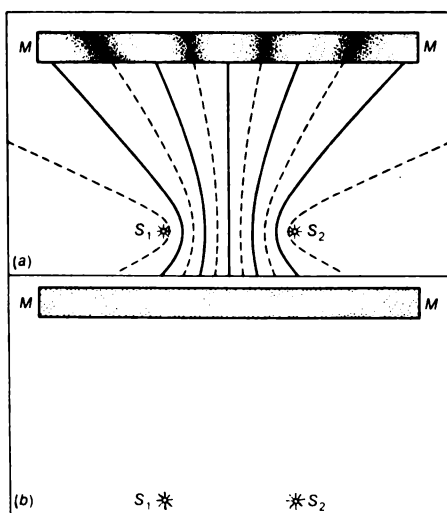
Arrangement of maxima and minima on the interference pattern.

We shall call the difference in distances from a point to sources  $S_1$  and  $S_2$  the *path difference* of two interfering waves to this point. Then the rule established above can be formulated as follows.

*The maxima of the interference pattern produced by two sources vibrating in phase are observed at points where the path difference is equal to an integral number of wavelengths (or, which is the same, to an even number of half-waves), and the minima are located at points where the path difference is equal to an odd number of half-waves.*

If two coherent sources  $S_1$  and  $S_2$  were not in phase but the wave from source  $S_2$  emerged by a fraction of a period later than the wave from  $S_1$ , it would easily be seen that the interference pattern, remaining stable, would be displaced towards  $S_2$ . Indeed, the waves would now meet in phase not at the points equidistant from the sources but at points such that a shorter time would be required for the wave from  $S_2$  to reach them. This time is equal to the difference between the time required for the wave from  $S_2$  to travel to a point equidistant from  $S_1$  and  $S_2$  and the time lag at the emergence of the wave. Then the rule formulated above should be changed accordingly.

Thus, when two coherent waves are superimposed, a stable interference pattern is formed, indicating that the energy of the waves is redistributed here: there are regions where the wave intensity is larger than the ordinary sum of the intensities of the two waves (maxima), but there are also regions where the intensity is smaller than the sum of the intensities of the two waves



**Fig. 93.**

- (a) Plate  $MM$  arranged in the path of coherent waves and intersecting the lines of maxima and minima is "illuminated" nonuniformly. In the regions of maxima, the wave intensity is larger than the sum of the intensities, while in the regions of minima, the intensity is smaller than this value.  
 (b) If the waves are incoherent, the plate is "illuminated" uniformly: the intensities are added.

(minima). If the total energy emitted by two sources remains unchanged, then the energy is just redistributed (Fig. 93*a*). On the other hand, when two incoherent waves are superimposed, the intensities are just added so that the addition of the second wave at each point leads to an increase in the wave intensity by the amount equal to the intensity of the second wave. Thus, no maxima or minima are observed (Fig. 93*b*).

#### 5.4. Interference of Acoustic Waves

Interference, like diffraction, is observed for any wave phenomena irrespective of the nature of waves. We analyzed main regularities of interference with the help of the waves propagating on the surface of water. Later, we shall consider the interference of electromagnetic waves used in radio engineering. In Part 3 of this volume devoted to physical optics, we shall analyze in detail the interference of light waves for which this phenomena was initially studied. The interference is also encountered in acoustics.

In order to observe the interference of acoustic waves, we must carry out an experiment similar to that with the waves on the surface of water (see Sec. 5.2). Two identical tuning forks emitting sounds in unison are fixed to a board that can be rotated about a vertical axis (Fig. 94). If the frequency of the tuning forks is about 1 kHz and their separation is 1.5 m, the width of alternating regions of enhanced and weakened sound, arranged in the horizontal phase in the same way as in Fig. 91, will be about 1 m (from maximum to maximum) at a distance of 5-6 m from the tuning forks.

If the tuning forks are excited (say, with a violin bow), and the board is slowly rotated, the regions of enhanced and suppressed sound will move relative to an observer who will hear loud sound and very weak sound alternately.

The experiment becomes even more illustrative if the observer hears only with one ear, the other ear being covered by palm. Besides, the room where this experiment is being carried out must be large and free of obstacles since waves reflected at them may strongly distort the interference pattern. In particular, the board with the tuning forks must be further from the floor and walls. If a vacuum-tube oscillator of acoustic frequencies is available, two telephones can be used instead of the tuning forks. The telephones should be connected in series to the oscillator and emit moderately loud sounds. Otherwise, if a too strong current is passed through them, they will emit non-

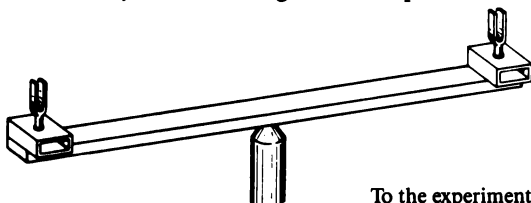


Fig. 94.

To the experiment with interference of acoustic waves.

sinusoidal oscillations, i.e. higher harmonics will become noticeable and blur maxima and minima of the sound. When a clear interference pattern is obtained from two telephones (or tuning forks), a check experiment can be made: having short-circuited a telephone (or silenced a tuning fork), one can verify that the interference pattern vanishes, i.e. there is no alternation of enhanced and weakened sounds.

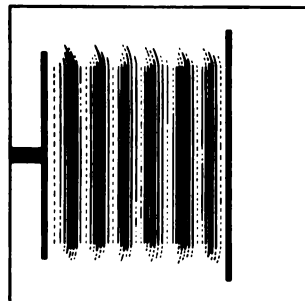
This experiment directly confirms that sound is a wave phenomenon. Moreover, if we know the distance between the sources of sound and measure the angle of rotation of the board between two adjacent minima of loudness (intensity), we can calculate the wavelength of sound in air. In general, interference is widely used for measuring wavelengths since the change in the path length of two waves from a minimum (maximum) to the next one is just equal to the wavelength.

### 5.5. Standing Waves

A special type of interference, known as a *standing wave*, is observed when *two coherent waves of the same intensity propagate against each other*. The superposition of such waves is always observed when a wave is incident on an obstacle with a good reflecting properties, arranged at right angles to the direction of wave propagation. Indeed, according to the law of reflection, the reflected wave will propagate in this case against the incident wave and will have the same intensity if the obstacle reflects the wave completely. The coherence of the forward and backward waves is ensured by the fact that they are just the parts of the same wave.

Let us make an appropriate experiment in the water bath.

A plate is placed on the path of a wave produced by a ruler disturbing water parallel to the ruler, i.e. at right angles to the direction of propagation of the wave (Fig. 95). The experiment reveals the following phenomena. When the wave running from the ruler is reflected and moves back, a series of stationary fringes emerges between the reflecting plate and the ruler parallel to them and separated by a half-wave. As always during interference, these fringes are alternating maxima and minima, the water surface at the minima being practically at rest.



**Fig. 95.**

Standing wave on the water surface.



Such is a standing wave on the water surface. Similar standing waves can also be obtained in a cord described in Sec. 4.2. In this section, we watched the propagation of the wave running from the hand of the experimenter along the cord till the moment when this wave reaches the point of suspension. But what will happen after that?

The wave is reflected at a fixed point of the cord and runs downwards along it, being superimposed on the forward wave produced by vibrations of the hand. Thus, here too a standing wave should be formed, and this is indeed the case.

Figure 96 shows the form of vibration of the cord. Stationary points on the cord alternate with the points where the vibrational amplitude is maximum. Stationary points are known as *nodes*, while the points where the amplitude has the maximum value are called *antinodes*. The distance between two adjacent nodes (or two adjacent antinodes) is equal to half a wavelength. The higher the frequency of vibration of the lower end of the cord, the shorter the wavelength, and the larger the number of nodes and antinodes formed in the cord. It is difficult to obtain many nodes and antinodes with the help of the hand since its movement should have a very high frequency. We can use instead a small electric motor rotating a simple crank mechanism. Arranging this mechanism horizontally and fixing the lower end of the cord to it, we can obtain a large number of nodes and antinodes as shown on the right-hand side of Fig. 96.

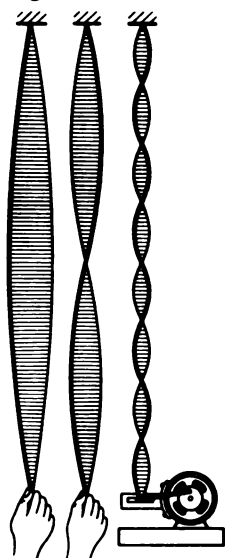


Fig. 96.  
Standing waves on a cord.

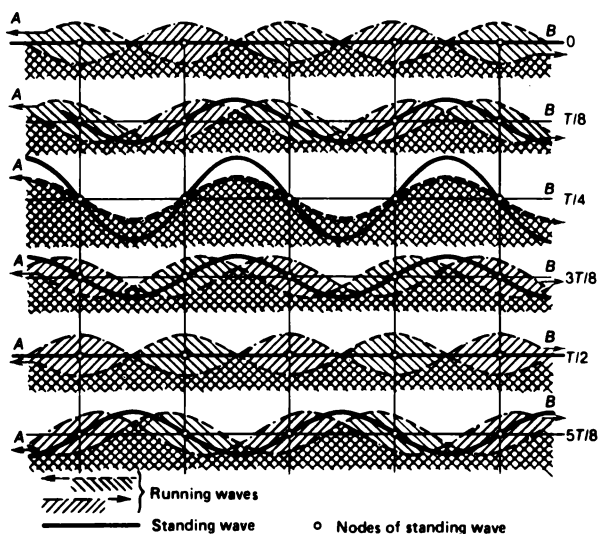


Fig. 97.  
Emergence of a standing wave as a result of  
superposition of similar opposite waves.

Why does the superposition of two opposite running waves result in alternating nodes and antinodes?

Figure 97 shows how it is done. The dashed and dot-and-dash lines in this figure indicate two waves propagating against each other. The figures arranged one below another give an idea of the pattern of the process after every one eighth of a period. During this time, running waves move towards each other along the straight line  $AB$  by one eighth of the wavelength. At each point of the straight line  $AB$ , the algebraic sum of the deviations from  $AB$  (with the plus sign corresponding to upward deviations and the minus sign to the downward deviations) is found. The points obtained in this way are connected with one another by a solid line. Thus, the solid line represents the result of summation of the two running waves.

If we watch the behaviour of the solid line from figure to figure, it can be seen that it always passes through the equilibrium positions at points marked by light circles. In other words, there is no vibration at these points which are hence the nodes of the standing wave. On the contrary, at antinodes formed between the nodes, the amplitude is at its maximum. *All points lying between two adjacent nodes vibrate in the same phase. But as we go over from one internodal gap to the next one, the phase changes by  $180^\circ$ .*

### 5.6. Vibrations of Elastic Bodies as Standing Waves

Each of the two similar running waves forming a standing wave transports an energy along the direction of its propagation. Since these directions are opposite to each other, *there is no energy transport in a standing wave. The energy "at rest", being converted from the kinetic to the potential energy and back* (this is the main ground for calling such a wave "standing"). Thus, the process occurring in this case is the same as that in elastic vibrations considered earlier, say, during the vibrations of a tuning fork or a plate fixed at one end in a vice. In both cases, we deal with harmonic vibrations of particles of a body, occurring at a certain frequency which is determined by the size and the properties of the given body. Moreover, different parts of the same body vibrate with different amplitudes. True, in the case of a vibrating plate, we observed only one point that remained at rest (the "node" was at the fixed end of the plate). On the other hand, during vibrations of a cord, a large number of nodes can be observed. It will be shown in subsequent sections, however, that both the tuning fork and the plate can be set in vibration at a higher frequency so that several nodes are also formed on them.

Thus, there is no difference between elastic vibrations of a body and standing waves in it: *vibrations of elastic bodies are just standing waves in them.*

While obtaining standing waves in a cord, we sustained these waves externally by the movements of the hand or by the crank mechanism. In other

words, these were forced vibrations with a frequency thrust by our action and equal to the frequency of this action. However, standing waves can also be free. By striking a tuning fork, a bell or an ordinary tumbler, drawing back and then releasing an elastic plate or a stretched string, we excite vibrations which are just free standing waves. Naturally, such vibrations gradually attenuate due to friction and other losses.

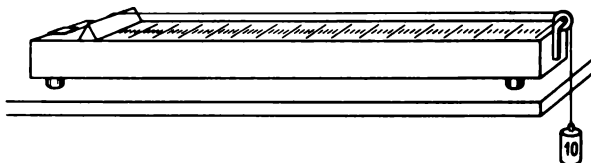
Let us now consider free standing waves using, by way of an example, the vibrations of a stretched string for which such waves can easily be obtained and observed.

### 5.7. Free Vibrations of a String

Experiments with a string can be conveniently carried out with the help of a device shown in Fig. 98. One end of the string is fixed while the other is passed over a pulley so that a load can be suspended on it. Thus, we know the tension of the string: it is equal to the weight of the load. The board above which the string is stretched is supplied with a scale. This allows one to rapidly determine the length of the entire string or of a part of it.

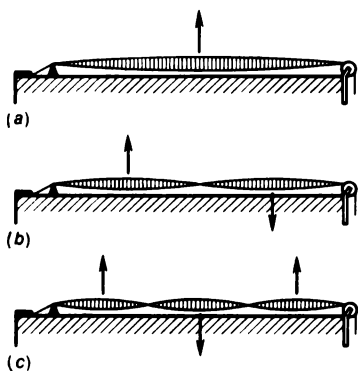
Pulling back the string at the middle and releasing it, we excite its vibration shown in Fig. 99a. Nodes will be at the ends of the string and an antinode, at its middle.<sup>1</sup>

Using this set-up and varying the mass of the load and the length of the string (by shifting an additional clip relative to the fixed end), it is not difficult to establish experimentally the parameter determining the natural



**Fig. 98.**

A device for investigating the vibrations of a string.



**Fig. 99.**

Free vibrations of a string: (a) with a single antinode, (b) with two antinodes, and (c) with three antinodes.

<sup>1</sup> Such a form of vibrations does not appear immediately, but sets in very soon.

frequency of vibrations of the string. The frequency  $\nu$  of string vibrations turns out to be directly proportional to the square root of the tension  $F$  of the string and inversely proportional to its length  $l$ :

$$\nu = k \frac{\sqrt{F}}{l}.$$

As to the proportionality factor  $k$ , it was found to depend only on density  $\rho$  of the material of the string and on the thickness  $d$  of the string. Namely,  $k = 1/d\sqrt{\pi\rho}$ . Thus, the natural frequency<sup>2</sup> of the string is given by

$$\nu = \frac{1}{ld} \sqrt{\frac{F}{\pi\rho}}.$$

In string instruments, the tension  $F$  is naturally created not by suspending the loads but by stretching a string by winding one of its ends over a rod (peg). The string is tuned to a required frequency by rotating the peg, i.e. by varying tension  $F$ .

Let us now proceed as follows. We pull one half of the string upwards and the other half downwards so that the midpoint of the string is not displaced. Releasing the two points (separated from the ends of the string by a quarter of its length) of the string simultaneously, we see that a vibration excited in the string will have, in addition to two nodes at the ends, one more node (at the middle), and hence two antinodes (Fig. 99b). For such a free vibration, the sound produced by the string will be twice higher (higher by an octave, as it is put in acoustics) in comparison with the previous vibration with a single antinode. In other words, the frequency is now  $2\nu$ . The string is as if divided into two shorter strings with the previous tension.

Further, we can excite vibrations with two intermediate nodes, which divide the string into three equal parts, to obtain vibrations with three antinodes (Fig. 99c). For this purpose, the string should be pulled at three points as is shown in Fig. 99c by arrows. The frequency of this vibration will be  $3\nu$ . By pulling the string at several points, we can in principle obtain vibrations with a larger number of nodes and antinodes, although it is rather difficult to realize. Such vibrations can be excited, for example, by touching the string with a bow at a point where an antinode should be obtained and slightly pressing with fingers at the points where the nearest nodes should be. Such free vibrations with four, five, etc. antinodes have the frequencies of  $4\nu$ ,  $5\nu$ , ....

Thus, a string has a set of vibrations, and hence a set of eigenfrequencies multiple to the lowest frequency  $\nu$ . The frequency  $\nu$  is known as the *fun-*

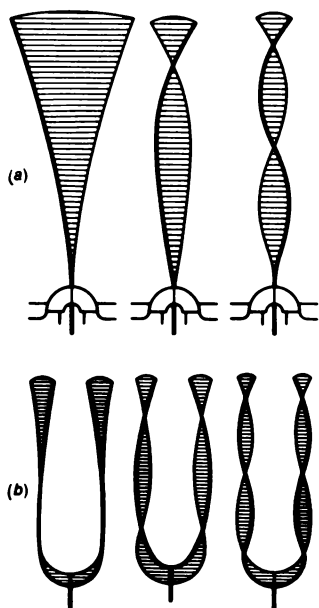
<sup>2</sup> If damping is low, it almost does not affect the frequency of free oscillations (see Sec. 1.11). Therefore, we always deal with the natural frequency, viz. the frequency of idealized completely undamped oscillations.

*damental frequency*, the vibration with frequency  $\nu$  is called the *fundamental tone*, while vibrations with the frequencies  $2\nu$ ,  $3\nu$ , etc. are called *higher harmonics* (*overtones*).

In string musical instruments, the vibrations of a string are excited either by plucking or jerking with a plate (guitar or mandoline), by knocking with a hammer (piano), or by a bow (violine or chello). The strings vibrate not with one of the eigenfrequencies but with several eigenfrequencies simultaneously. This is one of the reasons behind different timbres of different musical instruments (see Sec. 2.4) since overtones (higher harmonics) accompanying the fundamental vibration of a string are manifested differently in different instruments. (Other causes of different timbres are associated with the structure of the casing of an instrument, viz. its shape, size, rigidity, and so on.)

The presence of the entire set of natural vibrations and the corresponding set of eigenfrequencies is typical of any elastic body. However, in contrast to vibrations of a string, the frequencies of overtones are generally not always integral multiples of the fundamental frequency.

Figure 100 shows schematically the vibrations of a plate fixed in a vice at one end and of a tuning fork at the fundamental frequency and at two nearest overtones. Naturally, nodes are always observed at fixed points, while the maximum amplitudes are always obtained at free ends. The higher the overtone, the larger the number of additional nodes.



**Fig. 100.**

Free vibrations at the fundamental frequency and the first two overtones: (a) of a plate fixed in a vice, and (b) of a tuning fork.

When we earlier mentioned a single eigenfrequency of elastic oscillations of a body, we always meant the fundamental frequency and just ignored the existence of higher harmonics. However, when we considered vibrations of a load on springs or torsional vibrations of a disc on a wire, i.e. elastic vibrations of a system in which the entire mass is practically concentrated in one region (a load or a disc) while deformation and elastic forces are concentrated in another region (string or wire), we had all grounds for isolating the fundamental frequency. As a matter of fact, the frequencies of higher harmonics, starting from the second one, are much higher than the fundamental frequency. For this reason, overtones are practically not manifested in experiments with a fundamental vibration.

### 5.8. Standing Waves in Plates and Other Extended Bodies

Standing waves may form in bodies of any shape and not only in highly extended bodies like a string or a cord. The stationary regions of a standing wave, viz. its nodes, are the surfaces cutting the volume of a body into regions with the strongest vibrations (antinodes) located at their middle.

Strictly speaking, we have nodal surfaces, viz. stationary cross sections, in the case of a string or cord as well. But since the size of these sections is very small in comparison with the length of the cord or string, we speak of nodal points, treating the bodies as geometrical lines.

If a body has a shape approaching a geometrical surface, i.e. is a plate (flat or bent) or a shell, the nodal surfaces on this body can be treated as nodal lines. Figure 101 shows the vibration of a tumbler being struck at the edge. The solid lines represent nodal lines, while the dashed lines show (in an exaggerated form) how the tumbler walls are bent in this (fundamental) vibration. The bell vibrates in a similar way.

A visual and beautiful method of observing standing waves in plates was invented in 1787 by the German physicist Ernst Chladni (1756-1827). Sand was sprinkled on a glass, metal or wooden plate fixed at a point. Standing waves in the plate are excited by a bow rubbed with colophony, which is

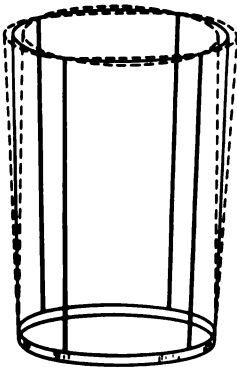


Fig. 101.

Vibration of a tumbler (fundamental vibration).

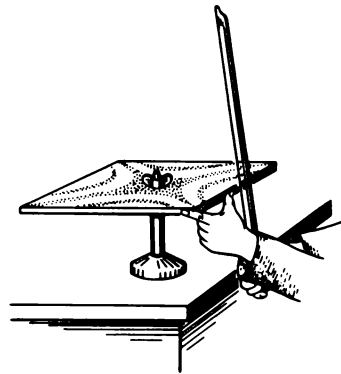


Fig. 102.

Obtaining Chladni's figures.

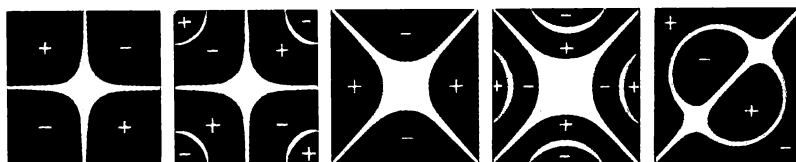


Fig. 103.

Examples of Chladni's figures. The plus sign marks antinodes where the plate is bent up at a given instant, while the minus sign corresponds to bending down. In a quarter of a period, the plate becomes flat, and after the next quarter, the plus will correspond to bending down and the minus, to bending up.

drawn across an edge of the plate (Fig. 102). Sand is shaken off the antinodes and is accumulated at nodal lines, forming the so-called *Chladni figures* which represent the pattern of nodal lines formed on the surface of the plate during its vibration. The shape of the figures depends on the shape of the plate and on the position of the fixed point as well as on the positions of the point at which the bow touches the plate and the point where the plate is held by fingers. Figure 103 shows Chladni's figures on a rectangular plate.

As an example of standing waves in the bulk of a body is the vibration of air contained in a solid shell (which is not necessarily closed). Let us take a rectangular wooden box open at face  $A'B'C'D'$  (Fig. 104). If the air vibrates along edge  $AA'$ , we obtain for the fundamental vibration (corresponding to the lowest frequency and the maximum wavelength) a nodal plane on wall  $ABCD$  and an antinode at opening  $A'B'C'D'$ . Thus, a quarter of a wave fits into length  $AA'$  of the box (Fig. 105a). For the second harmonic (first overtone), we have two nodal surfaces: one, as before, is on face  $ABCD$ , where a node should be formed in any case, and the other at a dis-

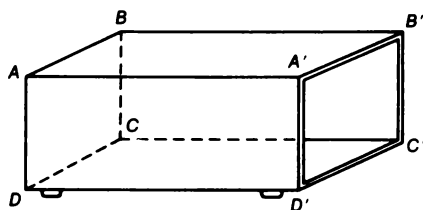


Fig. 104.  
A box open at one face.

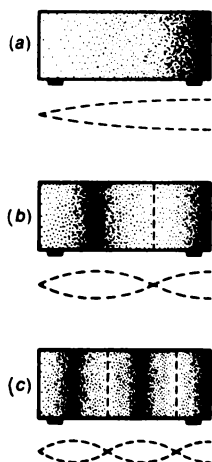


Fig. 105.

Standing waves in the box shown in Fig. 104: (a) fundamental vibration, (b) first overtone, (c) second overtone.

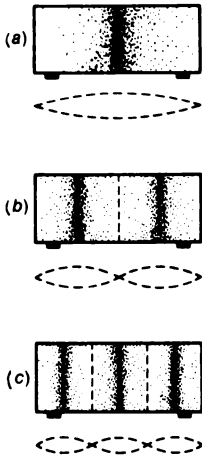


Fig. 106.

Standing waves in a closed box: (a) fundamental vibration, (b) first overtone, (c) second overtone.

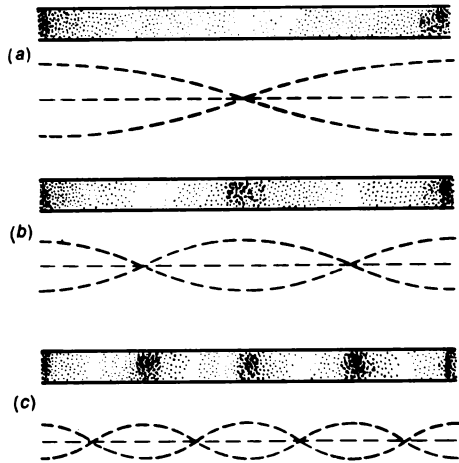


Fig. 107.

Standing waves in a pipe open at both ends: (a) fundamental vibration, (b) first overtone, (c) second overtone.

tance of a half-wave from this wall and a quarter of a wave from the open end at which we again have an antinode. Now  $3/4$  of a wave fits into the edge  $AA'$  (Fig. 105b), i.e. the wavelength is equal to one third of the fundamental wave, while the frequency is thrice as high as the fundamental frequency. The frequency of the second overtone (third harmonic) is five times as high as the fundamental frequency (Fig. 105c), and so on.

If we close the box, the nodal plane must be formed both on  $ABCD$  and  $A'B'C'D'$  for any natural vibrations along the edge  $AA'$ . Figure 106 shows the fundamental vibration and the first two overtones for this case.

The same type of standing waves is observed in pipes of various cross sections. Figure 107 represents the fundamental vibration and first two overtones in a circular pipe open at both ends. Antinodes are formed at both ends here.

Vibrations of air columns in tubes are used in wind musical instruments (like organ or flute).

### 5.9. Resonance in the Presence of Many Frequencies

It is well known that resonance phenomena, i.e. an abrupt increase in the amplitude of forced vibrations of a system, are observed when the frequency of the force coincides with the natural frequency of a system. What will be the situation when a system has a set of natural frequencies instead of one?

Let us consider in greater detail the forced vibrations of a cord whose lower end is attached to a crank mechanism (see Fig. 96). The rotational fre-

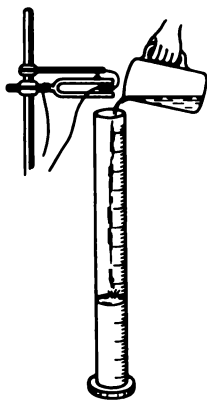


quency of this mechanism can be smoothly varied with the help of a sliding-contact rheostat connected to the circuit of the electric motor which drives the crank mechanism. Varying in this way the frequency of the force, we see that the most distinct nodes and the most bulging antinodes are formed on the cord when an integral number of antinodes exactly fit into the length of the cord, i.e. when the frequency of the force coincides with an eigenfrequency of the cord.

Thus, *if we have more than one eigenfrequency, resonance phenomena under the effect of a harmonic force are observed when the frequency of the force coincides with any eigenfrequency of the system.* All what was said above concerning a single eigenfrequency is applicable to any of these eigenfrequencies (see Sec. 1.13).

Of course, resonance phenomena can be obtained not only by varying the frequency of the force, but also by changing the eigenfrequencies of a system so that they in turn coincide with the frequency of the force, which remains unchanged. Let us arrange a sounding tuning fork above a tall cylindrical vessel having a height of about 50 cm (Fig. 108). The tuning fork with a sufficiently high eigenfrequency should be chosen for this experiment for the wavelength in air be not very large, e.g.  $\nu = 1000$  Hz ( $\lambda = 34$  cm). It is desirable to ensure undamped vibrations of the tuning fork using, for example, a breaker (see Fig. 56).

Pouring water into the vessel, we shall find that at certain levels of water, the sound of the tuning fork is amplified considerably. These are just the levels for which the height of the air column remaining in the vessel is equal to an odd number of quarters of the wavelength (Fig. 105). The frequency of the tuning fork consecutively coincides with the second overtone (third harmonic) of the air column when its height amounts to  $5\lambda/4$ , the first over-



**Fig. 108.**

Resonance of an air column to the sound of a tuning fork.

tone (second harmonic) at a height of the column of  $3\lambda/4$ , and with the fundamental frequency (when the height of the column is equal to  $\lambda/4$ ).

The sound is amplified during the resonance since intense vibrations of air above the area of the opening of the vessel produce a more intense acoustic wave in the surrounding air than that produced by the vibrating tuning fork itself (the reason behind this phenomenon will be considered in the following section).

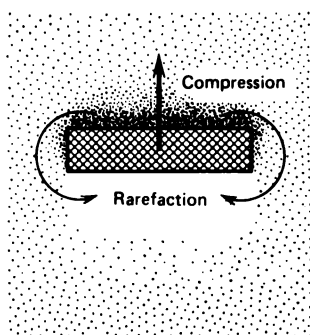
Therefore, when the sound of a tuning fork has to be amplified, it is fixed on a resonance box as was mentioned earlier (see Sec. 2.5, Fig. 40). When the tuning fork is sounding, its rod vibrates in the longitudinal direction. Being fixed to the upper surface of the box, it makes this surface bend upwards and downwards. As a result, air is alternately pushed out of the box and sucked into it. Thus, vibrations of the air column are excited in the box. The length of the box is chosen to be equal to a quarter of the wavelength produced by the tuning fork in air. Consequently, the fundamental frequency of the air column in the box open at one end is tuned in resonance with the frequency of the tuning fork. An intense vibration is excited in the box (see Fig. 105*a*), and the sound emitted from the opening is much stronger than the sound produced by the tuning fork itself.

The operation of the Helmholtz resonators mentioned in Sec. 2.7 is also based on the resonance of vibrations of air confined in the resonator cavity. From all the frequencies contained in an acoustic wave incident on the wide opening of a resonator (see Fig. 43), the latter responds most strongly to those which are equal to the eigenfrequencies of air vibrations in it. The most pronounced resonance is observed when the open cavity is in resonance with the fundamental frequency of air vibrations in it. The frequencies of higher harmonics are much higher than the fundamental frequency.

### 5.10. Conditions for a Perfect Sound Emission

In the previous section, it was pointed out that resonators considerably amplify the sound of tuning forks. Is this only due to the fact that the air column in a resonator is in resonance with the frequency of a tuning fork or are there some other conditions which are important here? Let us try to answer this question.

Let us consider the processes occurring near a prong of a sounding tuning fork. As the prong moves to either side, a condensation of air, and hence an increase in its pressure, is observed in front of it, and a rarefaction of air and a drop in pressure takes place behind it. As a result of this pressure difference, the levelling out of pressure (and density) occurs on both sides of the prong (Fig. 109). The process of pressure levelling out propagates at the same velocity as the acoustic wave, i.e. embraces a space with a radius equal to a half-wave during half a period. But the size of the prong of the tuning fork is much smaller than a half-wave. Therefore, compressions and rarefac-

**Fig. 109.**

The top view on the prong of a tuning fork. The thick arrow indicates the direction of its motion, while thin arrows show the direction of propagation of a compression wave around the prong.

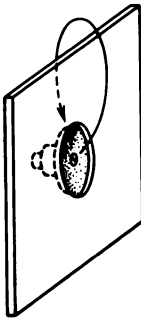
tions of air produced by it are considerably weakened due to levelling out of the pressure on both of its sides. This means that the emitted wave is also considerably weakened. We arrive at the conclusion that *for a good emission, the size of a body must not be small in comparison with the wavelength in the surrounding medium*. This explains the advantage of the resonance box since its length is equal to a quarter of a wavelength, and the levelling out of the air pressure around it is less pronounced than around the prong of the tuning fork.

We can easily draw another conclusion: a vibrating body emits higher frequencies (for which the wavelength is not large in comparison with the size of the body) better than low frequencies since the levelling out of the pressure has a stronger effect for longer waves. For example, the membrane of a dynamic loudspeaker whose diameter is about 15 cm emits frequencies which are higher than 2000 Hz well and is a poorer emitter of lower frequencies. This circumstance deteriorates the timbre of the sound. In order to improve it, the levelling out of the pressure on both sides of the membrane should be hampered for longer waves. For this purpose, the loudspeaker is fixed in a hole made in a wide board (Fig. 110), which increases the distance between the front and rear surfaces of the membrane. When such measures are taken, the emission of sounds at lower frequencies is considerably improved.

A tuning fork is a poor emitter not only because the wave emitted by each prong has a low intensity, but also because the prongs separated by a distance considerably shorter than the wavelength vibrate against each other, i.e. in antiphase. Therefore, the wave produced by a prong of the tuning fork at any point of surrounding air is suppressed due to the interference with the wave emitted by the other prong, which is in antiphase with the former wave.

Obviously, if we eliminate (or at least weaken) the vibration of one of the tuning fork prongs, we must gain in the intensity of sound. Indeed, it can easily be seen that if we enclose one of the prongs into a cardboard tube (Fig. 111), the sound will be enhanced.

How do the properties of the surrounding medium affect the emission of a given vibrating body?



**Fig. 110.**  
A loudspeaker mounted in a wide board.



**Fig. 111.**  
A tuning fork sounds stronger when one of the prongs is enclosed in a cardboard tube.

For a given frequency and amplitude of vibrations, the kinetic energy of the particles constituting a medium is the higher, the larger their mass, i.e. the higher the density of the medium. Under the same conditions, the potential (elastic) energy is the higher, the larger the rigidity of the medium, i.e. the lower its compressibility. Consequently, *for a given frequency and amplitude of vibrations of a source, it produces the wave whose intensity is the higher, the higher the density and elasticity of the medium.* For example, a plate vibrating in water emits a wave whose intensity is a few thousand times higher than the intensity of the wave emitted by the same plate vibrating in air.

### 5.11. Binaural Phase Effect. Sound Direction Finding

Let us consider again running waves propagating in air and analyze some phenomena determined by the arrangement of the source of these waves.

If the source of a sound is just in front of an observer or behind him, every compression or rarefaction of air in the acoustic wave reaches simultaneously both ears (Fig. 112a). Consequently, oscillations of air pressure are in this case in phase in both ears. If, however, the source is displaced to the right (or to the left), the waves first reach the right (left) ear (Fig. 112b), and oscillations of air pressure in the ears proceed with a phase shift.

The intensity of sound is practically the same in the two ears since the difference in the distances from the source is insignificant and the size of the head is not large enough to create a noticeable "sound shadow". In other words, if we neglect very high frequency, it can be said that acoustic waves round the head well enough (due to diffraction). Thus, the difference in vibrations in the two ears mainly consists in the phase difference between them.

It turns out that it is just due to the phase shift between the vibrations in the two ears that we can determine the direction to the source of the sound. This phenomenon is known as the *binaural phase effect*.

If an observer applies telephone earphones to his ears, and the sound of the same frequency is supplied to both telephones so that the phase shift between the right and left telephones can be artificially varied (this can easily be done by using electric methods), it will seem to the observer that the direc-

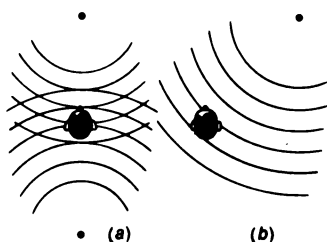


Fig. 112.

Phase difference of vibrations in two ears depends on the direction from which a wave arrives.

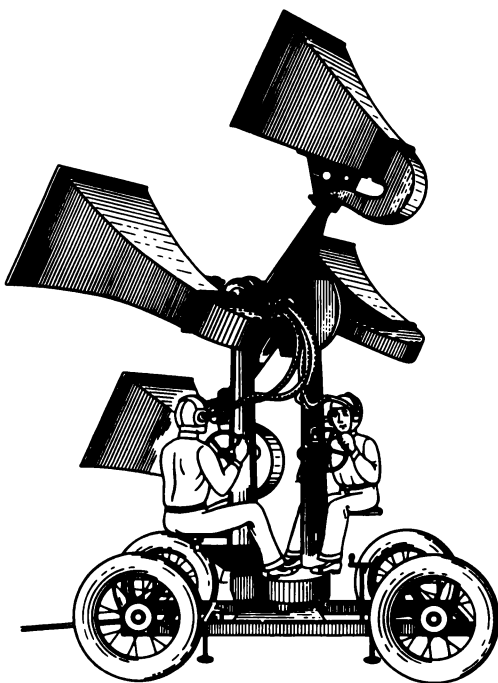


Fig. 113.

Sound locator.

tion to the source of sound changes. If the phase shift varies continuously, it will seem to the observer that the source of sound moves around him.

The binaural phase effect plays an important role not only in everyday life (we turn the head to the sound and find direction to sounding objects), but is also used in the so-called *sound direction finding*, viz. in determining the direction to a source of sound (an aeroplane, a battery, etc.). During World War II, specially trained soldiers (monitors) received the required sounds with the help of special sound locators (Fig. 113) and determined the direction to a source of these sounds.

Loud hailers (megaphones) are used not only for the amplification of sound. The direction to a source of sound is determined due to the phase difference in vibrations in two ears. When loud hailers are employed, this difference will be equal to the phase difference at their openings. Since the distance between these openings is much larger than between the ears, any deviation of the loud hailers from the direction to the source of sound will lead to a larger phase difference than the turn of the head through the same angle. Thus, loud hailers make sound direction finding more accurate.

## Chapter 6

# Electromagnetic Waves

### 6.1. Electromagnetic Waves

In Secs. 4.1. and 4.2, we have mentioned the electromagnetic waves and the extremely high velocity of 300 000 km/s at which they propagate in vacuum. It should be recalled and emphasized once again that electromagnetic waves emerge due to the *interrelation* between the variations of the electric and the magnetic fields. Any change in the electric field strength at a point in space induces a varying magnetic field at neighbouring points, whose variation in turn generates a varying electric field. It is for this reason that oscillations of the electric and magnetic fields are transferred from one point in space to neighbouring points, i.e. the electromagnetic wave propagates.

It is well known that an electric field is produced by electrically charged bodies, while a magnetic field embraces conductors through which an electric current flows (i.e. electric charges move). If electric charges are at rest, the electric field produced by them remains the same all the time (does not change). If charges move (e.g., in a metal wire) uniformly, we have a direct current which produces a constant, invariable magnetic field. Thus, in both cases the electric and magnetic fields are constant, and no electromagnetic wave can emerge.

On the other hand, for a nonuniform motion of electric charges, in particular, for any oscillation of charges, and hence for any alternating current, the electric and magnetic fields vary with time. These variations are transferred from point to point, and hence propagate in all directions, forming an electromagnetic wave.

It would seem that an electromagnetic wave can be obtained quite easily. For instance, we can make a charged body perform vibrational motion, or pass the alternating current of the lighting system through a wire coil. Since the electric field will vary in the former case and the magnetic field in the latter case, an electromagnetic wave should emerge in conformity with what had been said above. However, no wave phenomena acceptable for observation will take place in actual practice.

What is the reason behind this failure?

In order to answer this question, we must consider in greater detail the emergence of electromagnetic waves and find out the condition under which they are emitted well enough.

## 6.2. Conditions for a Perfect Emission of Electromagnetic Waves

As was mentioned earlier, the mutual relation between the electric and magnetic fields is manifested in an electromagnetic wave: a variation of one of these fields generates the other field.

The emergence of an electric field as a result of a variation of a magnetic field is just the *electromagnetic induction* experimentally discovered in 1831 by M. Faraday (see Vol. 2, Chap. 15). The inverse phenomenon, viz. the emergence of a magnetic field as a result of any variation of an electric field was theoretically predicted by the English physicist J.C. Maxwell (1831-1879). Proceeding from the assumption on the existence of this phenomenon, Maxwell drew the conclusion that electromagnetic waves must necessarily emerge for any variation of the electromagnetic field.

Maxwell's theoretical hypothesis required an experimental verification. If an experiment proves the existence of such electromagnetic waves, this will be the confirmation of the entire line of reasoning put forward by Maxwell, including the assumption on the emergence of a magnetic field as a result of variation of an electric field. For a successful experimental verification of the theory it is required that observed phenomena be sufficiently intensive.

According to Maxwell's theory, the magnetic induction of the field generated by a variation of an electric field is the larger, the higher the rate of variation of the electric field. The situation is similar to that in electromagnetic induction where the strength of the electric field induced by a variation of a magnetic field is the larger, the higher the rate of variation of the magnetic field (see Vol. 2, Sec. 15.4).

Thus, a necessary condition for the generation of high-intensity electromagnetic waves is a sufficiently high frequency of electric oscillations. The frequency of the current of a lighting system (50 Hz) is insufficient for successful experiments. Much higher frequencies of electric oscillations are required.

It is well known that such frequencies, attaining  $10^{10}$  Hz and even higher, can be obtained in electric oscillatory circuits (see Sec. 3.2.). However, it would be not an easy problem to detect electromagnetic waves in experiments with such circuits.

As a matter of fact, a high frequency of electric oscillations in a circuit, which is a necessary condition for obtaining high-intensity electromagnetic fields, is insufficient for the radiation of electromagnetic waves by this circuit.

This is so since an oscillatory circuit is almost a closed circuit whose size is small in comparison with the wavelength corresponding to the natural frequency of the oscillatory circuit. In such a circuit, we can find for each element with a certain direction of current or the sign of the charge another

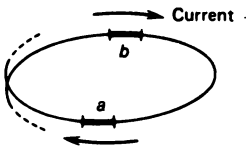


Fig. 114.

A turn of an inductance coil emits poorly since the elements with opposite currents are close to one another.

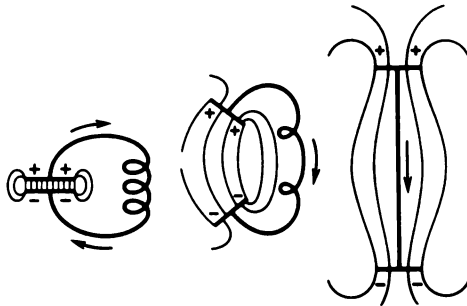


Fig. 115.

A transition from an oscillatory circuit to an open oscillator.

close <sup>1</sup> element with the opposite direction of current or unlike charge at the same instant. Let us take, for example, a turn of an inductance coil (Fig. 114). In any two diametrically opposite elements *a* and *b* of the turn, the currents are opposite to each other at any instant. Consequently, at large distances from the turn, elements *a* and *b* operate as two close antiphase emitters. The waves radiated by these emitters suppress each other everywhere like in the radiation emitted by the two prongs of a tuning fork (see Sec. 5.10). Since the turn as a whole is composed of such pairs of antiphase emitters, it is a poor emitter, which means that the entire coil is a poor emitter as well.

A similar situation takes place for the capacitor of the circuit: at any instant, the charges on the two plates are equal and opposite, these unlike charges being separated by a distance much shorter than the half-wave.

It follows hence what properties an oscillatory circuit must have to be a good emitter: we must go over to an open oscillatory circuit containing no elements with antiphase oscillations, or the distance between such elements must be comparable with the wavelength  $\lambda$ .

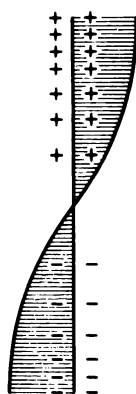
Figure 115 illustrates a transition from a nearly closed oscillatory circuit (interrupted by a thin insulating layer between the capacitor plates) to an open system known as *electric oscillator*. It is the simplest emitter of electromagnetic waves.

### 6.3. Oscillator and Aerials

In an open oscillatory circuit (oscillator), the charges are located not only on the capacitor plates but on the entire conducting wire of the oscillator. The presence of plates (flat or spherical) is generally not necessary. An oscillator can be made in the form of a straight wire. Having charged the oscillator so that the charges are distributed nonuniformly over its length, we create an electric field between separate elements of the oscillator. Under the action

<sup>1</sup> I.e. removed to a distance much shorter than a half wavelength.





**Fig. 116.**  
The charge density on an oscillator is indicated by the concentration of signs “+” or “-” and, besides, by the length of segments plotted perpendicularly to the oscillator (plus to the right and minus to the left).



**Fig. 117.**  
Current in an oscillator attains its maximum value at the midpoint and vanishes at the ends.

of this field, the charges start to move, and electric oscillations emerge. The methods of such a nonuniform charging of an oscillator will be considered later (Sec. 6.4).

During electric oscillations, the charges are accumulated with the maximum density at the ends of the oscillator, while the density of charges is always equal to zero at its midpoint (Fig. 116). With such a nonuniform distribution of charges, the oscillator cannot be characterized by a capacitance  $C$  concentrated in an element which is small in comparison with the wavelength generated by the oscillator as it could be done in the case of an oscillatory circuit with an ordinary capacitor.

The current is also different in different sections of the oscillator. When the charges flow from one half of the oscillator to the other, they are naturally accumulated at the oscillator ends so that the current at these ends is always equal to zero. In the middle region of the oscillator, the current is maximum (Fig. 117). Such a circuit in which the current is different in different cross sections of the wire cannot be characterized by an inductance  $L$  concentrated on a small region, as can be done for the oscillatory circuit with an inductance coil considered in Secs. 3.2. and 3.3. Thus, Thomson's formula determining the natural frequency of oscillations in a circuit is inapplicable to an oscillator. How can we find the natural frequency of electric oscillations in an oscillator? We shall use here the problem on the vibration of a string considered earlier.

It was shown that from the point of view of the theory of oscillations, swinging of a pendulum and electric oscillations in a circuit are similar phenomena (see Sec. 3.3). The difference is only in what does oscillate (simple oscillator in one case and charges in the oscillatory circuit in the other), but the law governing oscillations, i.e. the mode of oscillations, is the same in both cases. In this respect, the electric oscillations of a rectilinear oscillator are similar to the vibrations of a string or air column in a tube.

In the case of the string, we also could not use the formulas derived for vibrations of a simple oscillator. The mass of the string cannot be assumed concentrated in a small region (like the mass of the load in a simple oscillator), and the elasticity of the string concentrated in another region (like the spring of a simple oscillator). The mass of the string and its elasticity are distributed over its length. Similarly, the capacitance and the inductance are distributed over the entire length of an oscillator, unlike the Thomson oscillatory circuit, where the capacitance is concentrated in the capacitor and the inductance in the coil.

Accordingly, the laws governing the electric oscillations in an oscillator are similar to those governing elastic vibrations of the string. It can easily be seen that the current distribution in the oscillator (Fig. 117) is exactly the same as the amplitude distribution in a string fixed at two ends (Fig. 99*a*). The distribution of charge on the oscillator (Fig. 116) resembles the distribution of the amplitude of air column vibration in a tube open at two ends (Fig. 107*a*). Hence we may conclude that oscillations in an oscillator are just standing waves of current and charge. At the centre of the oscillator, there is the node of charge oscillations and the antinode of current oscillations, while at oscillator's ends there are nodes of current oscillations and antinodes of charge oscillations. Thus, a half-wave fits into the length of the oscillator, which is hence

$$l = \lambda/2.$$

But the wavelength of an electromagnetic wave is related to the oscillatory frequency through the formula  $\lambda = c/\nu$ , where  $c$  is the velocity of propagation of electromagnetic waves. Substituting this expression for  $\lambda$  into the previous formula, we obtain the following simple expression for the natural frequency of the oscillator:

$$\nu = c/2l.$$

This is the fundamental (lowest) natural frequency. Like in the string, oscillations at higher harmonics (overtones) can be observed in oscillators. Then two, three, four, etc. half-waves fit into the length of the oscillator. The frequencies of these higher harmonics are accordingly two, three, four, etc. times higher than  $\nu$ .

Figure 118 represents the time variation of current and charge. Figure 118*a* shows the oscillator at the instant of time when unlike charges on its two halves are maximal. At this moment, the electric field near the oscillator attains the maximum value, and the magnetic field is absent since there is no current. From this moment, the charges start to move from  $+$  to  $-$ , i.e. a current discharging the oscillator emerges (Fig. 118*b*). The current increases (together with the magnetic field) and attains its maximum value in a quarter of a period. By this moment, the oscillator is discharged complete-

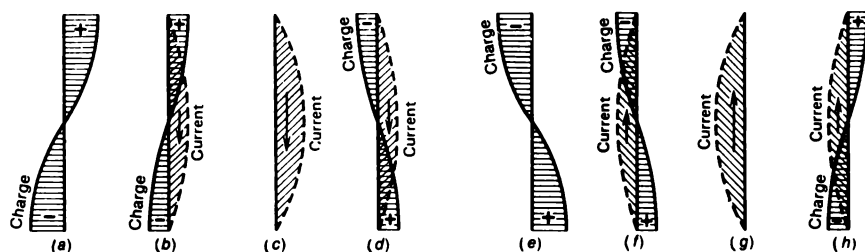


Fig. 118.

Oscillations of charge and current in an oscillator.

ly, and there is no electric field in the surrounding space (Fig. 118c). Flowing in the same direction downwards in the figure, the current recharges the oscillator: the positive charge is now accumulated at its lower part and the negative charge in the upper part (Fig. 118d). The current gradually attenuates and vanishes again by the end of the second quarter of a period. At this moment, the current (and the magnetic field) is zero again, while the charges (and the electric field) attain the maximum values with the opposite sign. This means that the oscillator is recharged (Fig. 118e). During the next half-period, the process described above is repeated with the opposite direction of the current (Fig. 118f-h). As a result, the initial state is restored by the end of the period (cf. Fig. 118a).

Thus, oscillations of charge and current in an oscillator proceed in the same way as those in an oscillatory circuit (see Sec. 3.2). The only difference is that in the case of the oscillatory circuit, the electric field (and hence the electric energy) can be assumed to be concentrated in the capacitor, and the magnetic field (and the magnetic energy) in the coil, while for the oscillator the electric and magnetic fields are distributed in the space around it. The same is observed as we go over from vibrations of a simple oscillator to those of a string: the potential energy of the simple oscillator is concentrated in the deformed spring, while the kinetic energy is concentrated in the vibrating load. On the other hand, both types of energy are distributed over the entire string.

It can be seen that at any instant of time, the current in the oscillator, having different magnitudes at different points, has the same direction. There are no regions with antiphase oscillations of the current. Further, charge oscillations in both halves of the oscillator are in antiphase (since the charges are unlike), but now the ends of the oscillator with charge antinodes are no longer close to each other but are separated by a distance equal to a half-wave. It is for this reason that an oscillator (and in general an open circuit, viz. an aerial) emits electromagnetic waves much better than an oscillatory circuit.

Hence it is clear why any modern radio transmitter contains, besides a continuous-wave oscillator, a certain open wire circuit known as aerial (radio antenna). The aerial is just the emitter of waves, which plays the same role as the resonator for a tuning fork or the sounding-board for a string instrument. Depending on the duty of a radio transmitter, the oscillator circuit diagram, the output power of the oscillator, the wavelength, the construction of the aerial, etc., may be different, but essentially, any transmitter contains a continuous-wave oscillator coupled with an open emitting circuit, viz. aerial (Secs. 6.7 and 6.8).

The energy emitted by an aerial is proportional to the power of electric oscillations in it, i.e. to the squared amplitude of these oscillations. For this reason, it is natural to try and increase the amplitude of oscillations in the aerial by tuning it in resonance with the oscillator frequency. For a simple oscillator, it is sufficient to make its length equal to a half-wave corresponding to the frequency of a CW-oscillator. However, this method can be used only as long as we deal with not very long waves. For wavelengths of tens of metres and more, aerials have a length shorter than a half-wave, and their tuning to resonance is realized by including an additional inductance coil. This coil can also be used for coupling the aerial with a CW-oscillator (Fig. 119). The earthing of the lower end of the aerial is equivalent to a (two-fold) increase in its length. For this reason, aerials intended for the reception of waves longer than one metre are normally earthed.

By varying the shape of an aerial, it is possible to obtain a directional radiation. For example, a simple vertical aerial emits uniformly in all horizontal directions (Fig. 120). An aerial, consisting of two vertical wires oscillating in phase and separated by a half-wave, strongly emits (due to interference) in the directions perpendicular to the plane of the wires (Fig. 121) and practically does not emit in the plane containing them.

#### 6.4. Hertz' Experiments on Electromagnetic Waves. Lebedev's Experiments

Maxwell's theory not only predicted the existence of electromagnetic waves but also indicated the conditions that are necessary for successful experi-

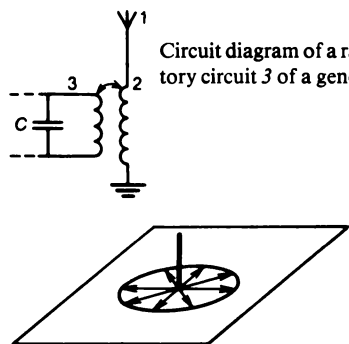


Fig. 120.

A simple vertical aerial emits uniformly in all directions.

Fig. 119.

Circuit diagram of a radio transmitter: aerial 1 is inductively coupled with oscillatory circuit 3 of a generator through loading coil 2. The lower end of the aerial is earthed.

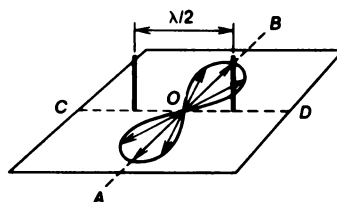


Fig. 121.

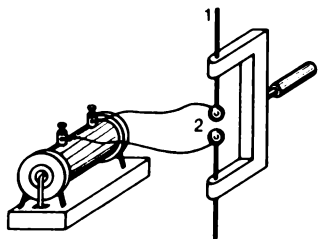
A two-wire broadside aerial radiates strongly in the directions of  $OA$  and  $OB$  and does not emit along  $OC$  and  $OD$ .

ments on their detection: a sufficiently high frequency of electric oscillations and an open oscillatory circuit. In 1888, Hertz carried out his famous experiments, trying to fulfil these conditions: he replaced an oscillatory circuit by a straight oscillator.

At his time, the only known method of exciting electric oscillations was the spark discharge. Figure 122 shows the schematic diagram of his set-up (Hertz' oscillator). Oscillator 1 is interrupted at the middle to form a spark gap 2. The ends of the oscillator are connected to a step-up transformer. The circuit diagram is quite similar to the one shown in Fig. 51, but now instead of a closed oscillatory  $LC$ -circuit we have an open oscillatory circuit which provides a good emission. The oscillations in this circuit are excited in the same way as described in Sec. 3.3 so that repeated outbursts of high-frequency damped oscillations emerge in the oscillator (see Fig. 52). The period of these oscillations, and hence the wavelength of electromagnetic waves, are determined by the size of the oscillator (see Sec. 6.3).

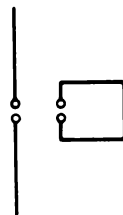
For wave detection, Hertz used another oscillator with a much shorter spark gap (fractions of a millimetre instead of 7.5 mm in the emitting oscillator). In addition to this receiving oscillator, a receiving loop made of a wire and having a rectangular shape and a short spark gap was also used (Fig. 123). Under the action of an electromagnetic wave, forced oscillations are induced in these receivers. If the receivers (an oscillator or a loop) are tuned in resonance with the emitter frequency, small and weak sparks are observed in their spark gaps under certain conditions which will be considered later (Sec. 6.6). Analyzing the appearance or absence of such sparks under various conditions for the emission and propagation of electromagnetic waves, and also for different arrangements of the receivers, one can draw a conclusion about the properties of the observed waves. In order to emphasize the difficulties in staging such experiments, it is sufficient to say that sparks in receivers can be detected only in a dark room by unfatigued eyes.

In his experiments, Hertz realized the reception of electromagnetic waves and managed to observe for these waves all phenomena typical of any wave process: the formation of a "shadow" behind well-reflecting (metal) objects, reflection at metal sheets, refraction in a large prism made of asphalt, and the formation of a standing wave as a result of the interference of a wave inci-



**Fig. 122.**  
Schematic diagram of Hertz' oscillator.

**Fig. 123.**  
Receiving oscillator and loop for Hertz' experiments.



dent at right angles to a metal sheet with the wave reflected by this sheet in the opposite direction. He also investigated the direction of vectors  $\mathbf{E}$  and  $\mathbf{B}$  of the electric and magnetic fields in electromagnetic waves. It turned out that electromagnetic waves possess the same properties as those of light waves (like polarization, see Sec. 6.6).

Thus, Hertz' experiments laid the solid foundation for the Maxwell theory: electromagnetic waves predicted by Maxwell's theory (see Sec. 6.2) were realized in experiment.

Further advances in the investigation of electromagnetic waves are associated with the name of the Russian physicist P.N. Lebedev (1866-1912). In 1895, he obtained (with the help of oscillators of the millimetre size) the waves with a wavelength of 6 mm which, as he wrote, were closer to the long waves of the thermal spectrum<sup>2</sup> than to the electric waves used initially by Hertz. For these waves, Lebedev observed all "optical" phenomena, viz. interference, polarization, reflection, refraction and even birefringence in a



Fig. 124.

Lebedev's devices for experiments with 6-mm electromagnetic waves.

<sup>2</sup> I.e. infrared waves.

prism cut out of crystalline sulphur. All devices for his experiments Lebedev made personally. The receiving oscillator consisting of two pieces of wire 3 mm long with a microscopic thermocouple soldered between them is a remarkable example of experimental art. Some original devices made by Lebedev are shown in Fig. 124.

### 6.5. Electromagnetic Theory of Light. Scale of Electromagnetic Waves

The theory of electromagnetic waves made it possible to interpret a large variety of electromagnetic phenomena from a unified point of view. However, this theory has led to one more conclusion of fundamental importance.

Using the results of measurement of purely electric quantities (like the forces of interaction between currents or charges), Maxwell managed to calculate the velocity at which electromagnetic waves must propagate. The result turned out to be striking: the obtained velocity was equal to 300 000 km/s, i.e. it *coincided with the velocity of light measured by optical methods*. Then Maxwell put forward his daring and revolutionary assumption that light in its nature is an electromagnetic phenomenon and that light waves are a kind of electromagnetic waves, viz. the waves with very high frequencies of the order of  $10^{15}$  Hz.

Hertz' experiments which confirmed the existence of electromagnetic waves and provided a verification for Maxwell's hypothesis that these waves propagate with the same velocity as light were a strong argument in favour of the electromagnetic theory of light. A large number of other phenomena, both known before and those discovered later, demonstrated such a close relation between optical and electromagnetic phenomena that the electromagnetic theory of light has become a well established fact rather than a hypothesis.

Investigations carried out in various branches of physics made it possible to establish that the range of frequencies or wavelengths<sup>3</sup> of electromagnetic waves is extremely wide. In this chapter, we shall confine ourselves to electromagnetic waves *in the narrow meaning of this term*, i.e. those with a wavelength larger than hundredths of a millimetre and mainly used in radio engineering (and hence called *radio waves*). Other (shorter) electromagnetic waves, their peculiar properties, the methods of their obtaining and observations will be considered in the following chapters. Here we shall present a diagram which gives an idea about the entire scale of electromagnetic waves.

This diagram (Fig. 125) has a peculiar structure due to a very large difference in wavelengths. The marks on the horizontal straight line separated by

---

<sup>3</sup> It should be recalled that frequency  $\nu$  is connected with wavelength  $\lambda$  through the relation  $\lambda = c/\nu$ , where  $c = 300\,000$  km/s.

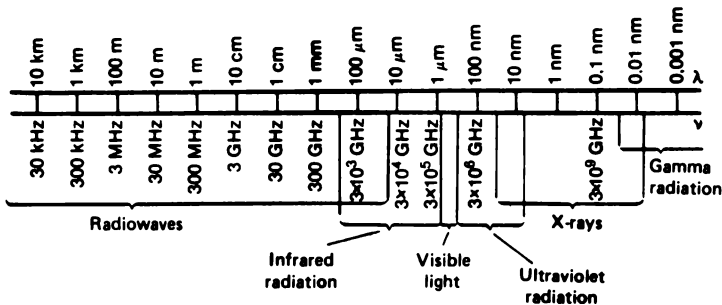


Fig. 125.

Scale of electromagnetic waves:  $1 \text{ GHz} = 10^3 \text{ MHz} = 10^9 \text{ Hz}$ ,  $1 \text{ nm} = 10^{-3} \mu\text{m} = 10^{-9} \text{ m}$ .

equal intervals correspond to wavelengths differing by a factor of ten from neighbouring wavelength. This is just the scale of wavelengths  $\lambda$  which starts in the diagram of the left with  $\lambda = 10 \text{ km}$  and finishes with the value  $\lambda = 0.001 \text{ nm}$ . Naturally, the values of  $10 \text{ km}$  on the left and  $0.001 \text{ nm}$  on the right are the limits of the diagram and not of the scale of electromagnetic waves, which can be extended in both directions.

The scale of wavelength  $\lambda$ , is followed by the scale of corresponding oscillatory frequencies  $\nu$ . Extending the scale to the left, we go over to longer waves, i.e. to lower frequencies until we reach the frequency  $\nu = 0$ , i.e. a constant (direct) current. It can be said that to such a current there corresponds an infinitely large wavelength, but this, of course, is a purely formal statement. As the frequency decreases, the emission conditions become worse and worse (see Sec. 6.2), and a direct current that should emit an “infinitely long” wave just emits nothing. Our diagram can also be extended to the right to higher frequencies and accordingly shorter waves.

The diagram presents the regions of  $\lambda$  (or  $\nu$ ) corresponding to various types of electromagnetic waves. As was mentioned above, in this chapter we shall limit ourselves to the left-hand region starting from “infinitely long” waves and finishing in the range of hundreds of micrometres, i.e. extending from the “zero frequency” to that of the order of  $10^{13} \text{ Hz}$ . It can be seen that the region of waves obtained by electric methods overlaps at its short-wave end with infrared (thermal) waves. This means that a wavelength of  $0.05 \text{ mm}$  can be obtained both from electric oscillations and by thermal methods, i.e. from a radiation of a heated body.

Not long ago, there were no overlapping regions on the scale of electromagnetic waves which, on the contrary, contained some gaps. In particular, there was a gap between the electromagnetic range (in the narrow sense) and infrared waves. Electromagnetic waves with a wavelength up to  $6 \text{ mm}$  (Lebedev) were obtained, while thermal waves had the limiting wavelength of  $0.343 \text{ mm}$  (Rubens).



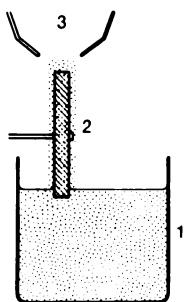


Fig. 126.  
Glagoleva-Arkad'eva's mass radiator.

In 1922, the Soviet physicist A.A. Glagoleva-Arkad'eva (1884-1945) filled this gap by obtaining electromagnetic waves with wavelengths ranging from 1 cm to 0.35 mm with the help of a device called a *mass radiator*.

The schematic diagram of this instrument is shown in Fig. 126. Vessel 1 contains small-size metal filings suspended in transformer oil. A stirrer which is not shown in the figure maintains the filings in the suspended state preventing them from falling to the bottom. A rotating wheel 2 entrains the mixture and is surrounded by it like by a tyre. A spark discharge is passed through the mixture with the help of wires 3 connected to an inductor. In their motion, metal filings form a large number of random couples which play the role of small oscillators emitting short waves during the discharge. Since the sizes of randomly formed oscillators are different and oscillations generated by them are not harmonic but rather damped, all wavelengths from the range indicated above are present in radiation simultaneously. It can be said that the mass radiator emits an "electromagnetic noise" rather than an "accord" or a "note".

In this device, two main difficulties, which are inevitable when a single oscillator of such a small size is used, have been overcome. Firstly, a single oscillator generates a negligibly weak radiation, while in a mass radiator a large number of oscillators operate simultaneously. Secondly, the filings in an ordinary oscillator rapidly burn out due to a spark. In the Glagoleva-Arkad'eva device this is not important since filings are continually replenished to the discharge region.

## 6.6. Experiments with Electromagnetic Waves

In order to reproduce some of the Hertz experiments and to get a better idea about electromagnetic waves, there is no need now to resort to an obsolete "spark" technique of excitation of waves. It is well known that the problem of obtaining continuous-wave electric oscillations has been solved with the help of self-induced oscillatory systems, viz. valve oscillators (see Secs. 3.5 and 3.6). It is essential that for a continuous-wave harmonic oscillation, the energy emitted by a transmitter is concentrated at a single frequency and not distributed over the entire spectrum as in the case of rapidly attenuating oscillations. For this reason, a receiver tuned in resonance with this frequency is in a much more advantageous conditions.

It is expedient to use for the experiments sufficiently short waves to be able to employ devices (resonance oscillators, screens, prisms and so on) of a medium size. Most convenient for this purpose are waves with a wavelength

of a few centimetres. Receivers and transmitters used at present at school laboratories mainly operate on the 3-cm waves.

Modern radio engineering also employs millimetre and even shorter (sub-millimetric) waves, but such short waves are not convenient for the experiments under consideration. These experiments can also be made with waves of the metre range (e.g., at a wavelength of 6 m when the length of a resonance oscillator is equal to 3 m). However, the centimetric and decimetric wave bands are more convenient since the experiments with 6-m waves should be made outdoors in an open flat place. Otherwise, the results are distorted due to the reflection of radio waves at surrounding objects (first of all, metal ones like iron bars in buildings, electric wiring, or telegraph transmission lines).

Let us describe some of possible experiments, assuming that a CW-oscillator is supplied with a radiating oscillator and a receiver, with a receiving oscillator.

**Reflected, refracted and standing waves.** In these experiments, the radiating and receiving oscillators should be arranged parallel to each other (e.g., both of them should be vertical).

When a CW-oscillator is switched on, the pointer of a galvanometer in the receiver is deflected. If a metal screen (like a sheet of metal), whose size is large in comparison with the wavelength (see Sec. 4.9), is arranged between the radiator and the receiver, a shadow formed behind it can be observed: the current in the galvanometer abruptly drops to zero when the oscillator

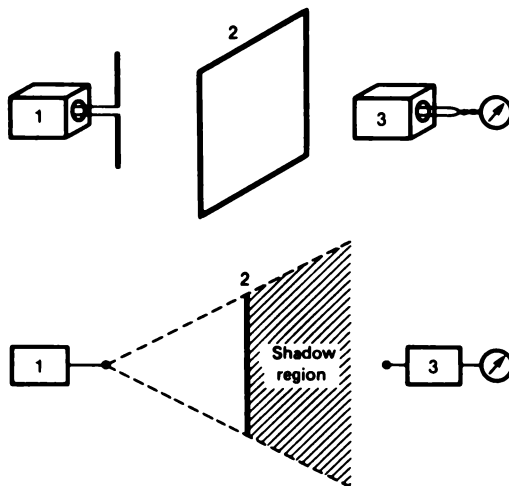


Fig. 127.

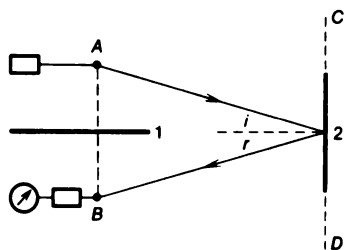
Formation of a shadow. The lower part of the figure shows the top view on the set-up: 1—generator with emitting oscillator, 2 — screen, 3 — receiver with an indicator.

is shielded by the screen. When the screen is removed or the receiving oscillator is taken out of the shadow region, the current increases again (Fig. 127). A human body also casts a noticeable shadow: if somebody passes between the radiating and the receiving oscillators, the current in the galvanometer first drops and then increases.

Taking a sheet of cardboard, thick wooden board or in general a screen of insulating material instead of the metal screen, we can easily see that such a screen is transparent for electromagnetic waves under consideration.

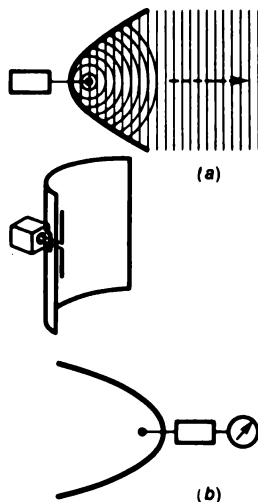
Having shielded the receiver from the radiator by a metal sheet 1 (Fig. 128), we can easily observe the reflection of electromagnetic waves at another metal sheet 2. Moving sheet 2 along the straight line  $CD$  parallel to segment  $AB$  (connecting the radiator with the receiver), we see that the strongest response (of indicator) is observed when sheet 2 is opposite to the midpoint of segment  $AB$  and its plane is parallel to  $AB$ . Thus, we verify in this way the validity of the law of reflection stating that the angle of incidence is equal to the angle of reflection (see Sec. 4.8). A replacement of the metal sheet 2 by a screen made of an insulating material shows that the reflection at such a screen is very weak.

The reflection by metals can be used for obtaining a directional radiation in the form of a nearly plane wave. For this purpose, an oscillator must be placed at the focus of a cylindrical mirror made of a metal sheet bent along a parabolic arc (Fig. 129a). The intensity of the plane wave emerging from such a reflector is considerably higher than that of nondirectional radiation of the oscillator in the absence of a reflector.<sup>4</sup> A similar reflector can be at-



**Fig. 128.**

Reflection of an electromagnetic wave:  
 $i$  — angle of incidence,  $r$  — angle of reflection.



**Fig. 129.**

A parabolic reflector of a radiating oscillator (a) and receiving oscillator (b).

<sup>4</sup> We mean here the lack of directionality in the plane normal to the oscillator.

tached to the receiving oscillator (Fig. 129*b*), which improves its sensitivity. The above experiments are therefore more illustrative when oscillators supplied with reflectors are used. The wires leading from the radiator to the oscillator are passed through the hole drilled in the reflector and having a size of 1-2 wavelengths. The wires leading from the receiver to the galvanometer are passed through small holes drilled in the reflector. The size of the reflector must be 3-5 times larger than  $\lambda$ .

The following experiment shows that an electromagnetic wave passing from one transparent material to another undergoes refraction, i.e. changes the direction of propagation. The refraction of waves at the boundary between two substances is one of the general wave phenomena. It was not considered earlier in detail since it is not very easy to observe this phenomenon for acoustic waves and surface waves in water. (Refraction can be observed and studied best of all for light waves, and will be considered in detail in the part of the book devoted to geometrical optics.)

In order to observe experimentally the refraction of an electromagnetic wave with a wavelength, say, 3 cm, a prism with a refraction angle of  $30^\circ$  (Fig. 130) should be made of paraffin or asphalt. The size of this prism must be large in comparison with the wavelength. Figure 131 shows how the direction of propagation of the wave changes as a result of refraction in such a prism. In the absence of the prism the strongest response in the receiver is observed in position *A*, while in the presence of the prism the wave is refracted, and the maximum response is observed in position *B*. Refraction takes place at two faces of the prism: when the wave goes over from air to paraffin and then at its exit from paraffin to air. The angle of deflection of the wave from the original direction of propagation ranges (depending on the material of the prism and the wavelength) between  $15$  and  $20^\circ$ .

Figure 132 represents the schematic diagram of the experiment for obtaining a standing electromagnetic wave. A flat metal screen is arranged opposite to the reflector so that the reflected wave propagates against the incident wave. If we move a receiving oscillator along the path between the

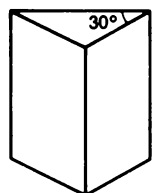


Fig. 130.

A prism made of paraffin or asphalt.

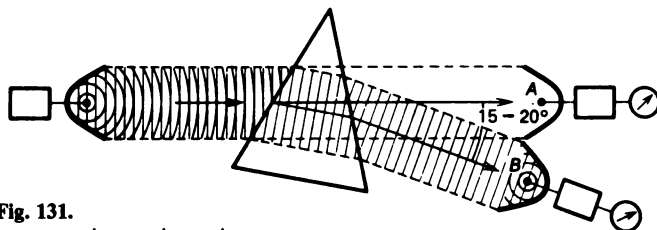
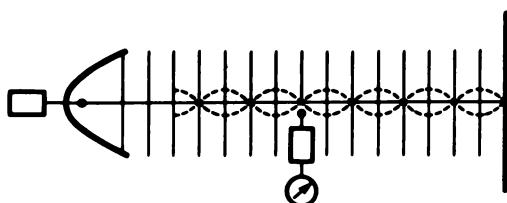


Fig. 131.

Refraction of an electromagnetic wave in a prism.



**Fig. 132.**  
Formation of a standing electromagnetic wave.

reflector and the screen, the current in the galvanometer will alternately increase (antinodes) and decrease (nodes).

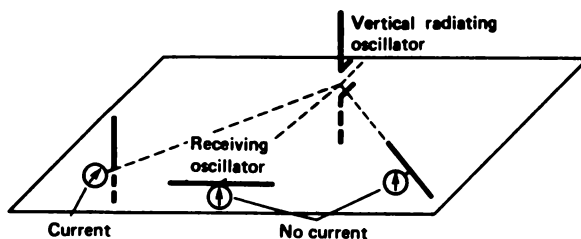
The separation between two adjacent antinodes or nodes is known to be equal to  $\lambda/2$  (see Sec. 5.5). If we know the frequency  $\nu$  of the oscillator beforehand, having measured  $\lambda$  by the method indicated above, we can determine the velocity  $c$  of propagation of the electromagnetic wave in air as follows:

$$c = \lambda\nu.$$

The most accurate measurements prove that this velocity coincides with the speed of light.

In the experiment described above, it remains unclear which antinodes and nodes are registered by the receiver (either the oscillations of electric or of magnetic field). The answer to this question will be given in the next chapter.

**Transverse nature of electromagnetic waves. Radio direction finding.** Standing at a fixed distance from a vertical radiating oscillator, let us transfer a receiving oscillator from the vertical to any horizontal position. It will be seen that the current in the indicator of the receiver drops to zero (Fig. 133). This can be explained only by the fact that the *electric field of the coming wave has vertical direction*. Indeed, such a field can move electric charges (cause a current) along the receiving oscillator only when the latter is in the vertical position and cannot do that when the oscillator is in horizontal posi-



**Fig. 133.**

The maximum current is induced in the indicator only for the vertical position of the receiving oscillator. The current is absent for any horizontal position of the oscillator.

tion. Hence it follows that in the above experiment with the standing wave, the receiving oscillator indicated the nodes and antinodes of the electric field.

Let us repeat the same experiment as that shown in Fig. 133, but now with a wire loop instead of the receiving oscillator. The result will be as follows. When the loop is arranged in the vertical plane, passing through the radiator, a current flows through it. But during any rotation of the loop through  $90^\circ$  from the indicated plane the current in it vanishes (Fig. 134). It is well known that the current in a loop (or coil) is induced by a varying magnetic field only if this field pierces the loop. Consequently, the absence of the current in the positions of the loop shown in Fig. 134 at the middle and on the right is explained by the fact that the *magnetic field of the incoming wave has the horizontal direction and is perpendicular to the direction of propagation*. Indeed, in this case the magnetic field pierces the loop in the first position and does not in the other two positions.

Thus, we arrive at the conclusion that the *electric field strength  $E$  and the magnetic induction  $B$  in the wave are at right angles to each other and to the direction of propagation of the wave* (Fig. 135). The direction of  $E$  coincides with the direction of oscillator, while vector  $B$  lies in the plane perpendicular to the oscillator.

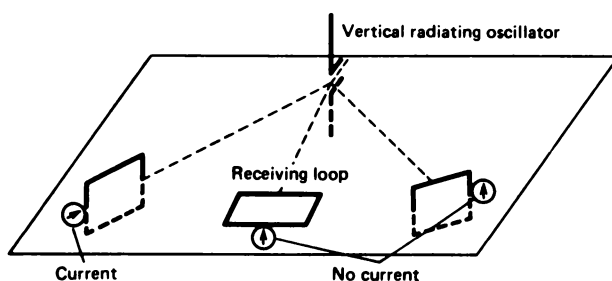


Fig. 134.

The maximum current in the receiving loop is observed when it is in the left position. The current is absent for the two other positions.

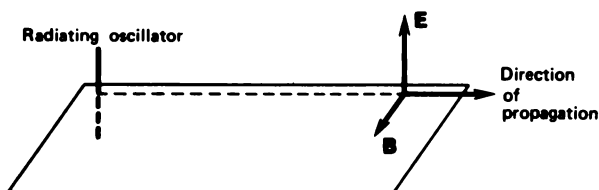


Fig. 135.

The arrangement of the electric and magnetic field vectors with the vertical position of an oscillator for waves propagating in the horizontal direction.

We analyzed the case of a vertical oscillator and the horizontal propagation of a wave. An analysis of other directions of propagation shows that a similar arrangement of vectors  $\mathbf{E}$  and  $\mathbf{B}$  is observed for any direction of propagation: (1) both vectors are perpendicular to the direction of propagation, and hence oscillate in the plane perpendicular to this direction, i.e. electromagnetic waves are transverse; (2) vector  $\mathbf{E}$  lies in the planes passing through radiating oscillator, while vector  $\mathbf{B}$  is normal to these planes (Fig. 136).

The *transverse nature of oscillations is a general property for any electromagnetic wave*, which depends neither on the choice of the direction of propagation nor on the origin of radiator. Another *general property is that vectors  $\mathbf{E}$  and  $\mathbf{B}$  are mutually perpendicular in an electromagnetic wave*. We shall consider this question again while studying light waves.

Returning to Fig. 136, it should be observed that if the directions of electric and magnetic fields  $\mathbf{E}$  and  $\mathbf{B}$  has been established, the direction of propagation of the wave is thus determined. In other words, we know the direction from the point at which the wave is received to the radiator. For almost all the aerials employed in engineering, the direction of the electric field is vertical. The direction of the magnetic field can be determined with the help of a receiving loop (or coil consisting of a few loops, viz. a frame, or loop antenna). This forms the basis of *radio direction finding*, viz. the determining of the direction from a given point to a radio station being received.

Figure 137 shows a movable radio direction finder, viz. a receiver equipped with a frame antenna which can be rotated about its vertical axis. Such an antenna can easily be made at home. Having connected it to an ordi-

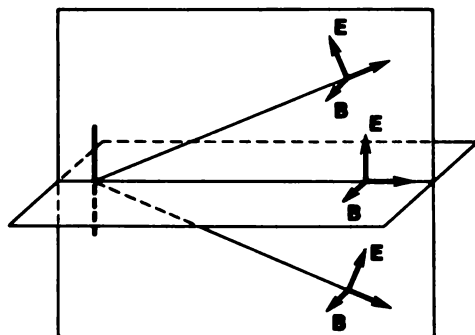


Fig. 136.

Transverse nature of electromagnetic waves.

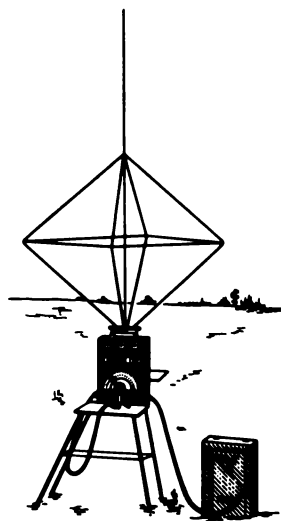
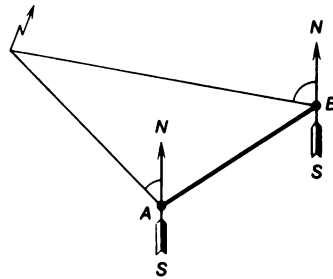


Fig. 137.

The appearance of a movable radio direction finder.



**Fig. 138.**  
Pinpointing of a radio transmitter  
from two points.

nary tube receiver (to the terminals marked “antenna” and “earth”), one can take bearings of a powerful radio station.

Normally, the frame antenna of a radio direction finder is arranged in a position for which the intensity of the reception passes through zero (this method is more precise than that based on the maximum intensity). For such a position, the magnetic induction  $\mathbf{B}$  of the wave lies in the plane of the antenna, and hence the direction to the radio station coincides with the straight line perpendicular to the antenna plane. The instrument does not indicate to which side of the antenna the station is pinpointed, but normally this is known beforehand.

If the direction to a radio station (bearing) is determined from two points the distance between which is known (points  $A$  and  $B$  in Fig. 138), plotting a triangle from the known side  $AB$  and two angles, one can pinpoint the radio station, i.e. determine its location.

The principle on which the direction finding is based is also used in *radio navigation*, viz. navigation and piloting along a certain direction indicated by special radio transmitters (radio beacons). A special receiver with a frame antenna (*radio compass*) is installed on a ship or aeroplane and indicates deviations from a required course. Sometimes the signals received by a radio compass are used for monitoring steering gears, i.e. for automatic maintaining the required course (autopilot).

### 6.7. Popov's Invention of Radio

It was mentioned above that Maxwell's theory was confirmed in experiments with electromagnetic waves. Hertz' experiments soon became known to scientists all over the world. The idea about the utilization of electromagnetic waves for communication and even for the wireless transmission of energy thus emerged. However, nobody could indicate practical ways of realization of this idea. Hertz himself apparently was not sure that the waves observed in his experiments could be used for communication because of their extremely weak effect. Such was the state of the art by the time when the Russian physicist and electrical engineer A.S. Popov (1859-1905) started his work. Having repeated Hertz' experiments, Popov improved the design of the set-



up and within a year (1889) managed to make the sparks in his receiving resonators visible in a large auditorium without darkening. Soon he became aware that for practical utilization of electromagnetic waves, a sensitive and convenient *receiver* is required first of all.

Such a receiver was designed by Popov by 1894. The main parts of its construction can be found also in modern receiving equipment. What were the basic features of the Popov's first receiver and what was its operation principle?

In order to improve the sensitivity of the receiver, Popov used the phenomenon of resonance. A significant merit of Popov's invention lies in the introduction of a *highly elevated receiving aerial* which considerably extends the range of reception and is still used in all radio receiving stations.

Another distinctive feature of Popov's receiver lies in the method of registering waves. For this purpose, Popov used instead of spark a special instrument, viz. *coherer* invented recently by Branly and used for laboratory experiments. The construction of a coherer is as follows. Fine metal filings are placed in a glass tube. Two wires are sealed into both ends of the tube so that they are in contact with the filings. Under normal conditions, the electric resistance between separate filings is comparatively large so that the coherer as a whole has a high resistance. An electromagnetic wave creates a high-frequency alternating current in the coherer circuit and causes tiny sparks between filings, which weld filings together. As a result, the resistance of the coherer abruptly drops. In order to increase the resistance of the coherer to the initial value and make it sensitive to electromagnetic waves again, it should be shaken. Popov introduced a coherer into a circuit containing a battery and a telegraph relay (Fig. 139). Before the arrival of an electromagnetic wave, the resistance of the coherer is high, the current flowing through it and through the relay is weak, and the armature is not attracted to the lower electromagnet. When an electromagnetic wave appears, the resistance of the coherer drops, the current increases sharply and the relay armature is at-

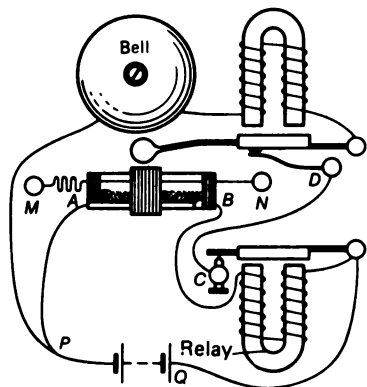


Fig. 139.

Schematic diagram of Popov's radio receiver from the paper published in Journal of Russian Physico-Chemical Society (January 1896).

tracted to the electromagnet. Thus, contact *C* connecting an ordinary electric bell to the battery is made. The hammer strikes the bell (or records a spike on a strip of moving paper), thus signalling the arrival of the wave. On moving backwards, the hammer strikes the coherer, thus restoring its sensitivity.

Thus, Popov realized the so-called *relay switching circuit* (see Vol. 2, Sec. 18.12). The negligibly low energy of arriving waves is used not directly for the reception (say, for the emergence of a spark), but is employed to control the energy source feeding the registering instrument. In modern radio receivers, the coherer is replaced by electron valves, but the relay principle remains in force: an electron valve essentially operates just as a relay. Weak signals supplied to the valve control the power of current sources feeding this valve.

Besides, in his receiver Popov realized the principle of the *feedback* which is still widely used in radio engineering. The amplified signal at the receiver output (the circuit of the electric bell) automatically acts on the receiver input (the circuit of the coherer). The feedback (realized in the device considered above by the electromechanical method) is the main new in principle element in the Popov's invention.

On May 7, 1895, Popov demonstrated the operation of his receiver at the sitting of the Russian Physico-Chemical Society. This date is rightfully considered as the birthday of radio.

Significant advances have been made in the field of radio engineering during the comparatively short span of time that has elapsed since Popov's invention. Even during the first few years following the invention, a number of considerable modifications was introduced, some of which were also worked out by Popov. In particular, he included in the receiver an ordinary telegraph, and, as a result, the arrival of an electromagnetic wave was indicated not only by the bell but also by a mark on the telegraph tape.

In his further investigations carried out together with P.N. Rybkin, Popov realized the audioreception of signals. It turned out that if signals are too weak for actuating the coherer, poor contacts between filings operate as a detector (Sec. 6.8), and each signal is accompanied by a sound in a telephone connected to the coherer. This discovery made it possible to extend the range of radio communication.

The next important step in the development of radio was made soon after its invention and consisted in the improvement of the transmitter. The spark gap was taken out of the aerial to form a special oscillatory circuit, which served as a source of oscillations. The aerial coupled with this circuit now operated only as a radiator of waves.

The invention of electron valves by the American scientist Lui de Forrest in 1906, paving the way for the creation of the sources of undamped electric oscillations (see Secs. 3.6 and 6.6) was extremely important for the develop-

ment of radio. This invention made it possible to transmit by radio not only telegraph signals, but also sounds of speech, music, etc., i.e. to realize wireless communication and broadcasting.

### 6.8. Modern Radio Communication

If a transmitter emits an undamped sinusoidal wave, a harmonic oscillation will be induced in receiving aerial. Obviously, no signal can be transmitted in this way. With the help of a receiver, we can only say whether or not the transmitter operates. In order to transmit certain signals, speech, music, television picture, etc., we must somehow change the nature of emission of the transmitter, for example, change the amplitude of its oscillations. This process is known as *modulation*. The simplest way of telegraph modulation consists in the interruption of radiation with the help of a key, i.e. in the transmission of short and long signals, viz. dots and dashes of the Morse code (Fig. 140). In telephone modulation, the amplitude of radiation is changed not by switching it on or off, but smoothly with sound frequencies being transmitted (Fig. 141).

Figure 142 shows a circuit diagram explaining the process of telephone modulation. In contrast to Fig. 58, the secondary of a step-up transformer, containing in his primary circuit an ordinary telephone inset (carbon microphone) and a battery, is connected between the grid and the cathode of a tube. Under the action of acoustic waves incident on the membrane of the

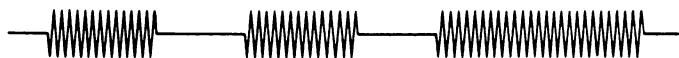


Fig. 140.  
Telegraph (pulse-type) modulation.



Fig. 141.  
Telephone (speech) modulation.

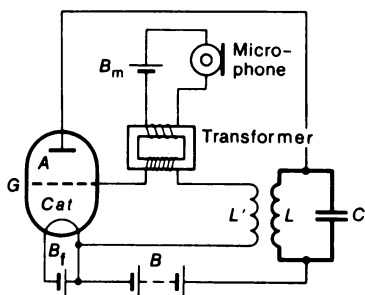


Fig. 142.  
Circuit diagram of telephone modulation

microphone, carbon powder contained in it experiences the pressure varying with the frequency of sound. As a result, the resistance of the microphone, and hence the current in the primary of the transformer, vary at the same frequency. This leads to the emergence of varying emf in the secondary of the transformer, i.e. a varying voltage of sonic frequency is applied to the grid of the tube. The amplitude of high-frequency oscillations generated in the LC-circuit by this tube varies with this low-frequency voltage on its grid. Therefore, the intensity of radio waves emitted by the antenna varies in the same way.

Naturally, modern radio transmitters are much more complicated, but the basic principles of their construction are the same as those described above.

The power of modern broadcasting stations attains many hundreds and even thousands of kilowatts. Special vacuum tubes whose size is sometimes larger than the height of a man are created for such stations.

The antenna of a receiver is acted upon by modulated radiation from a large number of simultaneously operating transmitting stations. Besides, various sources of noise (e.g., atmospheric electric discharges, sparkling contacts of electric machines and devices, and so on) generate electric oscillations in the receiving antenna. The aim of the receiver is (1) to *separate* oscillations of the required station from this mixture of oscillations; (2) to *amplify* the separated oscillations, and (3) to *obtain* from these high-frequency modulated oscillations the signals (oscillations at acoustic frequencies, telegraph and television signals, etc.) by which the radiation of the station have been modulated.

The first problem is known to be solved with the help of resonance (see Sec. 3.4). A receiver contains oscillatory circuits (or a single circuit in the simplest case) which single out a narrow frequency band (the so-called transmission band) from the entire complex set of electric oscillations in the antenna. By varying the tuning of oscillatory circuits of the receiver, we shift the transmission band along the frequency scale. To tune to a given radio station means to obtain a transmission band of the receiver, such that the frequency of the station falls into this band. Naturally, the transmission band also includes a fraction of oscillations from the sources of noise. The reception is possible only if oscillations from the station being received are not too weak in comparison with the level of noise oscillations.

The second problem, viz. the amplification of oscillations singled out with the help of resonance, is solved by using either vacuum tubes (see Vol. 2, Sec. 8.16) or semiconductor triodes (see Vol. 2, Sec. 9.3). Amplifying oscillations, these devices operate as relays: the gain in the intensity of oscillations is due to the energy of the sources (say, batteries) feeding a tube or a transistor. If a vacuum tube is used as an amplifier, weak oscillations of voltage generated by an electromagnetic wave in the oscillatory circuit are applied to the grid of the tube and cause much stronger oscillations in the anode circuit. Amplified oscillations from the anode of one tube can be supplied to the grid of the next tube and amplified still more (multistage amplification). At present, vacuum tubes are being replaced by semiconductor diodes and

triodes (see Vol. 2, Sec. 9.3) which are much smaller in size and require much lower feeding voltage and power.

Finally, the third problem, viz. the reconstruction of low-frequency modulating signals from high-frequency modulated oscillations, is solved with the help of demodulators (*detectors*), i.e. the *devices which conduct current in one direction better than in the opposite direction*. In modern radio receivers, vacuum tubes and semiconductor diodes (including point-contact diodes) are used as detectors. In the latter case, the rectifier is the contact between a semiconductor crystal and a cat's whisker. Rectifying contacts of this type (which, in addition, operate without any dc-source) were known as crystal detectors and were used in radio engineering even before vacuum tubes were invented. Let us consider the operation of such a detector.

In view of different resistance of a detector for two directions of current, the form (oscillogram) of the alternating current through the detector differs from that of the voltage supplied to it (Fig. 143). The voltage oscillations have equal deviations (amplitude) on both sides of zero (Fig. 143*a*), while current oscillations are "cut" on the side where the detector is poorer conductor (Fig. 143*b*). But such an asymmetric alternating current can be presented as the sum of a direct current (line 1 in Fig. 143*c*) and a symmetric alternating current (curve 2).

Thus, if a high-frequency sinusoidal voltage is applied to a detector, a direct current will pass through the detector in addition to a high-frequency current. This current can, for example, make deflect the pointer of galvanometer connected in series with the detector.

Let us now assume that the amplitude of the high-frequency voltage applied to the detector is not constant but modulated, viz. varies at a low frequency (Fig. 144*a*). In the detector, we obtain an asymmetric current which

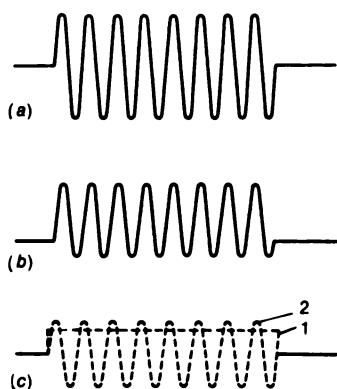


Fig. 143.  
Operation of a detector.

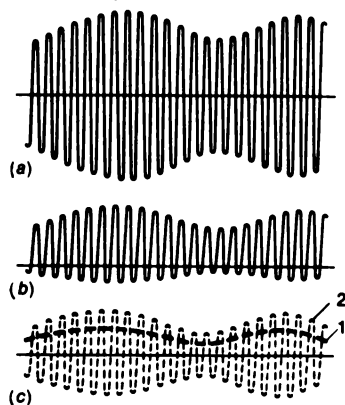


Fig. 144.  
Rectification of a modulated oscillation.

is also modulated (Fig. 144*b*). If we decompose this current, as it was done earlier, separating from it the high-frequency symmetric oscillation (curve 2, Fig. 143*c*), the second component will be not a direct current but a current varying with a low frequency, viz. the frequency of modulation (curve 1). If we connect a telephone in series with the detector, this low-frequency (acoustic) current makes the telephone membrane vibrate and hence can be heard. Such a simple combination of a detector and a telephone was used in the so-called crystal receiver (Fig. 145) which had been widely used before the receivers with vacuum tubes appeared.

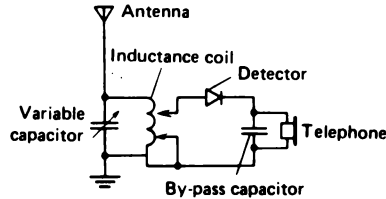


Fig. 145.

Circuit diagram of a crystal receiver.

A crystal receiver operates not on the relay principle, but uses directly the energy captured by the receiving antenna. The detector with a telephone is connected to a resonance oscillatory circuit so that the telephone is shunted by a by-pass capacitor through which the high-frequency component of the detector current can easily pass. The advantage of crystal receivers over tube receivers is the absence of feeding sources. However, this is also the cause of its main drawback due to which it was replaced by tube receivers, viz. its low sensitivity.

## 6.9. Other Applications of Radio

We mentioned above some other applications of radio besides radio communication and broadcasting. For example, it was shown that radio is used for determining the location of objects which do not emit radio waves but just reflect them (radio location, see Sec. 4.3), and also for determining the position of transmitters, radio beacons, etc. (radio direction finding, see Sec. 6.6). Let us now consider some other applications of radio.

Let us begin with the so-called *phototelegraphy* which allows one to obtain a photographic copy of an original (a drawing, photograph, or letter) located at a transmitting station. The principle of operation of a phototelegraph is very simple. At a transmitting station, there is a tube with a special device (photoelectric cell) fixed at one end and a special optical objective at the other end (Fig. 146). The objective of the tube is arranged above a roller over which an original to be transmitted is wound. The objective converges to the photocell only the light from a small region of the surface of the original, viz. the region above which the objective is located at a given moment. During operation, the roller rotates and translates parallel to its axis so that the objective consecutively (point by point) "inspects" the entire area of the original. The amount of light incident on the photocell through the objective varies in accordance with the alteration of light and dark regions on the original. The photocell has a property that the current through it depends on the intensity of the incident light (Sec. 21.2). Therefore, as the roller with the original moves, the current in the photocell varies in accord with the alteration of light and dark regions on the original. These variations of current are used for the modulation of

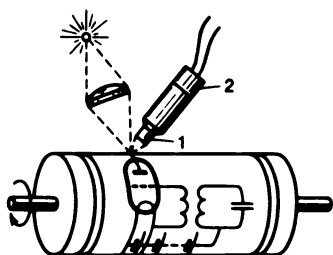


Fig. 146.  
Operation of transmitter phototelegraph: 1 — objective, 2 — photoelectric cell.

radio transmitter in completely the same way as the variations of current in the microphone circuits are used for modulation during the transmission of sound.<sup>5</sup>

At the receiving station, the low-frequency current repeating the oscillations of current in the photocell of the transmitter is obtained after the amplification and detection of the received radiation. If the current obtained after detecting feeds a small electric lamp, the brightness of the filament will vary in accord with the brightness of the points of the original which pass under the objective on the transmitting station.

At a receiving station, there is the same roller with an objective as at the transmitting station, but instead of the photocell, an electric lamp is fixed on the tube, and instead of the original, a sheet of photographic paper is wound over the roller. The objective forms a light spot on this paper, whose brightness varies with the intensity of light incident on the photocell of the transmitting station. If the roller at the receiving station moves in the same way as the roller at the transmitting station, a copy of the transmitted original will be obtained on the photographic paper after it has been exposed point by point and developed.

Phototelegraph operates at a comparatively low rate since its rapidity depends on the velocity of motion of mechanical appliances and on the rate of variation of the brightness of the lamp. For this reason, this method cannot be used for transmitting the images of moving objects (*television*).

In order to realize the principle of television, mechanical appliances and incandescent lamps had to be replaced by electron beams in the same way as mechanical relays suitable for telegraphy had to be replaced by vacuum tubes to make it possible to transmit sounds by radio (radio telephone).

The principle of television transmission is as follows. At a transmitting station, the roller moving under the photocell is replaced by an electron oscillograph in which an electron beam runs at a very high velocity not over a screen but over a complex multicell photoelectric device known as *iconoscope* (from the Greek words "iconos" meaning image and "scopio" meaning observation). The image to be transmitted is projected on this photocell with the help of an objective. Each cell of the iconoscope is actuated at the moments when an electron beam falls on it. Special scanning voltages applied to control plates of the oscillograph make the electron beam run over the entire surface of the iconoscope during 1/25 second (the end of the beam draws thereby 625 horizontal lines densely arranged one below another). The current in the circuit of the iconoscope at each instant is proportional to the illuminance of the cell on which the electron beam is incident at this moment. Therefore, oscillations of the current in the iconoscope circuit correspond to the light intensity distribution over the surface at all consecutively "inspected" points of the picture to be transmitted (frame).

Electric oscillations from the iconoscope are supplied to a radio transmitter and modulate the radio wave emitted by it in the same way as the alternating current in the microphone circuit modulates a radio wave during the sound transmission. Thus, every second a radio wave carries the "replicas" of 25 complete frames each of which consists of 625 lines.

<sup>5</sup> Phototelegrams can also be transmitted by wires. In this case, the varying current produced by a photocell is directly transmitted after amplification over the wires to the point of reception.

At the receiving station, the roller and the lamp of the phototelegraph are also replaced by a cathode-ray oscillograph, but now having an ordinary screen glowing under the impacts of electrons (the so-called kinescope). After the amplification and detection of the received wave, the same current that has modulated the wave in the transmitter is obtained in the receiver. This current is used to control the intensity of the electron beam in the kinescope. But the brightness of the glow of the kinescope screen is proportional to the intensity of the electron beam. Thus, the brightness of the spot on the screen of the receiver varies with time in accord with the illuminance of the points of the transmitted image through which the electron beam runs in the transmitter.

The electron beam in the receiver moves over the screen synchronously with the motion of the electron beam over the iconoscope of the transmitter, i.e. it runs over the entire area of the screen in  $1/25$  s, drawing thereby 625 horizontal lines. As a result, the entire frame is reproduced on the screen of the transmitter in  $1/25$  s. Since 25 such frames change one another per second, individual images are perceived by human eyes, like in cinema, as a single moving image.

By radio transmission of special signals (commands), say, certain combinations of telegraph characters, a control at a distance (*telecontrol*) can be realized. A remote receiver mounted, for example, on board a ship or aeroplane, Earth's satellite, etc. is tuned beforehand to the frequency of controlling transmitter. The receiver is either switched on permanently, or is switched on after certain periods of time automatically, according to a given program. Received and amplified signals-commands actuate certain relays which, in turn, switch on or off auxiliary electric motors fed by local sources of energy and performing various mechanical operations. Thus, powerful engines, steering gears, measuring instruments, radio transmitters, etc. can be controlled at a distance.

Let us mention one more application of radio.

It was shown in Secs. 5.2 and 5.3 that the waves running over the surface of water from two coherent sources form a typical interference pattern (see Figs. 91 and 92), consisting of stationary alternating lines of maximum and minimum intensity of oscillatory motion of the water surface. The Soviet scientists L.I. Mandelshtam and N.D. Papaleksi observed interference phenomena for radio waves and indicated their practical application. They employed phenomena for rapid and very accurate measurement of distances between different points on the surface of the Earth, thus creating a new branch of radio engineering, viz. *radiogeodesy*. The high speed of measurement makes it possible to use it when one of the points (a ship or aeroplane) moves. For this reason, this method of measuring distances can be used in practice for navigation and piloting by radio, viz. in *radionavigation*.

## 6.10. Propagation of Radio Waves

The laws of propagation of radio waves in free space are relatively simple, but in most cases in radioengineering one deals not with free space but with propagation of radio waves over the Earth's surface. Both theory and experiment show that the Earth's surface strongly affects the propagation of radio waves, so that not only the physical properties of the surface (like the difference between seas and land), but also its geometrical shape (the curvature of the globe and the difference in the relief of mountains, valleys, etc.) are important. These effects are different for different wavelengths and different separations between the transmitter and receiver.

The effect of the shape of the Earth's surface on the propagation of radio waves is explained by what has been said above. Indeed, we essentially deal with various manifestations of *diffraction* of waves emitted by a radiator (see Sec. 4.9) both on the globe as a whole and on elements of the relief. The diffraction is known to depend strongly on the relation between the wavelength and the size of a body lying on the path of the wave. Therefore, it becomes clear



that the curvature of the surface of the Earth and its relief differently affect the propagation of waves of different wavelengths.

For example, a mountain ridge casts a "radio shadow" for short waves, while sufficiently long waves (of the order of a kilometre) bend around this obstacle and are not weakened on the slope of the mountain behind a crest (Fig. 147).

As to the globe as a whole, it is too large even in comparison with the longest waves used in radio. Very short (metre) waves do not bend noticeably behind the horizon, i.e. beyond the range of sight. The longer the waves, the better they bend around the surface of the globe, but even the longest of the utilized waves cannot bend due to diffraction strongly enough to pass around the globe and reach our antipodes. Nevertheless, there exists radio communication between any points on the globe, but this is possible due to a completely different reason, which will be considered later, and not due to diffraction.

The influence of the properties of the Earth's surface on the propagation of radio waves is associated with the fact that under the influence of these waves, high-frequency electric currents are induced in the soil and sea water, which are the strongest in the vicinity of the transmitter aerial. A fraction of energy of a radio wave is spent for sustaining these currents which liberate certain amount of Joule's heat in the soil or in water. These energy losses (and hence the resulting weakening of the wave) depend, on the one hand, on the conductance of the soil, and on the other hand, on wavelength. Shorter waves attenuate much sooner than longer waves. For a good conductance of a medium (sea water), high-frequency currents penetrate to a smaller depth from the surface than for a lower conductance (soil), and energy losses in the former case are considerably smaller. As a result, the range of the same transmitter turns out to be (a few times) longer for waves propagating above the sea than for those propagating above the land.

It was noted above that the propagation of radio waves over very long distances cannot be explained by the diffraction around the globe. However, long-range radio communication (over a few thousand kilometres) was realized even in the first years after the invention of radio. At present, any radio amateur knows that long-wave ( $\lambda$  greater than 1 km) and medium-wave ( $\lambda$  ranging between 100 m and 1 km) stations can be heard during winter nights at a distance of many thousands of kilometres, while during daytime, especially in summer, these stations are heard only at distances of the order of a hundred kilometres. In the short-wave frequency range

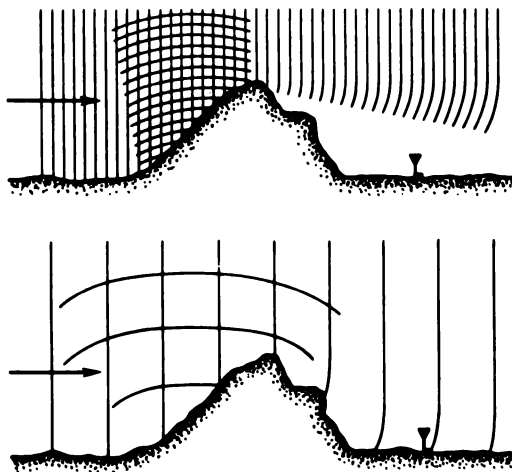


Fig. 147.

A mountain casts a "radio shadow" for short waves. Long waves bend around the mountain.

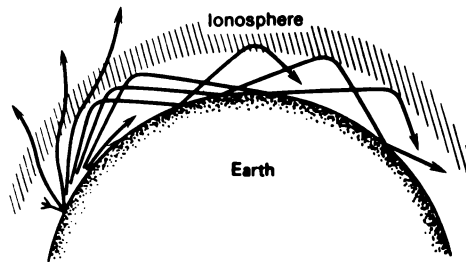


Fig. 148.

Propagation of a wave between the Earth and the ionosphere.

( $\lambda$  between 10 and 100 m), the situation is different. Here such wavelengths can be selected which reliably cover any distance in any time of the day and during any season. To ensure round-the-clock communication on such waves, one should operate at different wavelengths during different periods of the day. The dependence of the range of propagation of radio waves on the season and daytime indicated that there exists a relation between the conditions of propagation of radio waves on the Earth and the activity of the Sun. At present, this relation is well studied and explained.

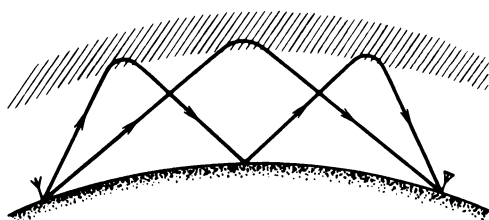
Along with visible rays, the Sun emits a high-intensity ultraviolet radiation and a large number of fast charged particles, which get into the Earth's atmosphere and strongly ionize its upper layers. As a result, a few layers of ionized gases lying at different altitudes (100 and 200-300 km) are formed. Because of the existence of these layers the term *ionosphere* has been applied to the upper layers of the atmosphere.

Due to the presence of ions and free electrons, the ionosphere has the properties which distinguish it from the rest of the atmosphere. Retaining the ability to transmit visible light, infrared radiation and metre radio waves, the ionosphere strongly reflects longer waves. For such waves (with  $\lambda$  greater than 10-15 m), the globe turns out to be enveloped as if by a spherical mirror, and these radiowaves propagate between two reflecting spherical surfaces, viz. the surface of the Earth and the "surface" of the ionosphere (Fig. 148). It is for this reason that radio waves can bend around the globe.

Naturally, the expression "the surface of the spherical mirror of the ionosphere" should not be understood literally. There is no sharp boundary between ionized layers, and the regular spherical shape is not observed either (at least, simultaneously over the entire globe). Ionization is different for different layers (it is higher in the upper layers and lower in the lower ones), and the layers consist of continually moving and changing "clouds". Such a nonhomogeneous "mirror" not only reflects, but also absorbs and scatters radio waves in different ways depending on the wavelength. Besides, the properties of the "mirror" vary with time. In the daytime, the ionization under the effect of solar radiation is much stronger than at night when positive ions and negative electrons just recombine to form neutral molecules (recombination). Especially sharp difference between the ionization during daytime and night is observed for the lower layers of the ionosphere. The air density here is higher, the collisions between ions and electrons are more frequent, and the recombination is more intense. At night, the ionization of the lower layers of the ionosphere may drop to zero. The ionization is different for different seasons, i.e. depending on the altitude of the Sun above the horizon.

An analysis of daily and seasonal variations of the state of the ionosphere has made it possible not only to explain but also to predict the conditions of propagation of radio waves having different wavelengths during different periods of the day or year (radio forecasts).

The existence of the ionosphere not only permits the short-wave communication over large distances, but also makes it possible for radio waves to bend around the globe, and even several times. As a result, a peculiar phenomenon is observed during radio reception, viz. *radio echo*,

**Fig. 149.**

Different paths of waves from a transmitter to a receiver.

due to which a signal is received several times: after the arrival of the signal along the shortest path from the transmitter, repeated signals bending around the globe can be heard.

It often happens that a wave moves from a transmitter to a receiver through several different paths and experiences different numbers of reflections from the ionosphere and the Earth's surface (Fig. 149). Obviously, the waves emitted by the same transmitter are coherent and can interfere at the point of reception, where they amplify or suppress one another, depending on the path difference. Since the ionosphere is not a perfectly stable "mirror" but rather varies with time, the path difference of the waves arriving along different paths from the transmitter to the receiver also changes with time. As a result, the amplification is changed by weakening, then again by amplification, and so on. The interference fringes can be said to "crawl" over the surface of the Earth, and the receiver is now at the maximum and then at the minimum of oscillations. In the program received, clear reception alternates with fading in the reception during which the audibility may fall to zero.

A similar phenomenon is sometimes observed on the screen of a TV set when an aeroplane flies above the receiving aerial. The radio wave reflected by the plane interferes with the wave from the transmitting station and the observed pattern "flickers" since the interference "fringes" of alternating amplification and weakening of the signal run (due to the motion of the plane) by the receiving aerial.

It should be noted that when a television program is received in a town, doubling (and even "multiplication") of the pattern on the screen of the kinescope is observed. The pattern consists of two or more images shifted relative to one another along the horizontal. This is the result of reflection of radio waves from buildings, towers, etc. Reflected waves traverse longer paths than the distance between the transmitting and receiving antennas, and hence lag behind one another, forming a pattern shifted in the direction of scanning of electron beam in the kinescope. We essentially observe the result of propagation of radio waves at a finite velocity of 300 000 km/s.

Since the ionosphere is transparent for radio waves with wavelengths shorter than 10 m, the radiation from extraterrestrial sources can be detected. This gave rise to *radio astronomy* which emerged as a science in the forties and is being rapidly developed. This science opened new prospects for investigating the Universe besides those available due to ordinary (optical) astronomy. The number of radio telescopes being constructed is on the increase, the size of their antennas grows, the sensitivity of receivers is being improved, and as a result the number of various extraterrestrial sources discovered with the help of radio astronomy continuously grows.

It turned out that radio waves are emitted by the Sun and planets, and beyond the limits of the solar system, by many nebulae and supernovas. A large number of sources of radio waves were discovered beyond our stellar system (Galaxy). They are mainly other galaxies, only a small fraction of which has been identified with optically observed nebulae. "Radio galaxies" were discovered at very large distances from the Earth (billions of light years), which are beyond the limits accessible for modern powerful optical telescopes. High-intensity sources of radio waves with very small angular dimensions (of the order of an angular second) have been discovered.

Initially, it was assumed that these were stars of special type belonging to our Galaxy, and hence they were called quasi-stellar sources, or quasars. But since 1962, it has become clear that quasars are extragalactic objects having a very high power of radio-frequency (r.f.) radiation.

Individual, or discrete radio sources of our Galaxy emit a broad spectrum of wavelengths. However, a "monochromatic" r.f. radiation at a wavelength  $\lambda = 21$  cm emitted by interstellar hydrogen has also been discovered. An analysis of this radiation made it possible to determine the total mass of interstellar hydrogen and to calculate its distribution over the Galaxy. Recently, a monochromatic r.f. radiation at wavelengths typical of other elements has also been discovered.

For all sources of r.f. radiation mentioned above, the intensity is strictly constant. Only in some cases (in particular, for the Sun) separate outbursts of r.f. radiation are observed against a background of a continuous radiation. The year of 1968 was marked by a new important discovery in radioastronomy: sources emitting strictly periodic pulses of radio waves were discovered (mainly within our Galaxy). These sources were called pulsars. The periods of repetition of pulses for different pulsars are different and range from a few seconds to a few hundredths of a second and smaller. The nature of r.f. radiation from pulsars is explained most plausibly by assuming that pulsars are rotating stars consisting mainly of neutrons (neutron stars). The large scientific value of this discovery for radioastronomy consists in the detection of such stars and in the possibility of their observation.

Besides the reception of the intrinsic r.f. radiation of the objects from the solar system, their radiolocation is also used. This is the subject of *radiolocation astronomy*. By receiving radio signals emitted by powerful radars and reflected from a planet, one can measure the distance to this planet very accurately, estimate the velocity of its revolution about an axis, and judge (from the intensity of reflection of radio waves of different wavelengths) about the properties of the surface and the atmosphere of the planet.

Concluding the section, it should be noted that the transparency of the ionosphere to sufficiently short radio waves makes it possible to realize all kinds of radio communication with the Earth's artificial satellites and spaceships (including, besides the communication as such, radio control, television and telemetry, viz. the transmission of readings of various instruments to the Earth). For this reason, it has become possible to use metric radio waves for communication and teletransmission between localities separated by large distances (say, between Moscow and Far East) using a single *relaying* of programs with the help of special repeater communication satellites.

### 6.11. Concluding Remarks

Radio has become one of the broadest and most important fields of engineering, which determines to a considerable extent the level of modern civilization. A new branch was formed in physics, known as *radiophysics* and dealing with the investigation of various phenomena pertaining to this field. While speaking of modern radio, it should be borne in mind that this term is no longer exhausted by the applications in which we deal with the propagation of radio waves over more or less large distances (radio communication, television, radiolocation, radio navigation, etc.). Quite different applications of radio engineering instruments and methods also play a very important role.

In various practical fields, the problem of conversion of nonelectric oscillations (mechanical and acoustic vibrations, oscillations of the intensity of light, temperature or pressure, the level of liquid, and so on) into electric os-

cillations often arises. This is associated with the fact that modern radio engineering methods allow one to readily and rapidly operate with electric oscillations, viz. amplify them a million times, vary their frequency and form, and finally observe and analyze them up to frequencies of  $10^8$  Hz with the help of cathode-ray oscillographs.

As a result, radio engineering instruments and methods of investigations have found their way into almost all branches of technology and are being widely used in various scientific investigations both in laboratories and in natural conditions (for instance, in the investigation of the ionosphere). They are also employed in industry and medicine. High-frequency electric oscillations are used for curing patients, for hardening steels, drying wood, sterilizing preserves, mine detecting, etc. A cathode-ray oscillograph is now a common instrument in an optical laboratory as well as on a biologist's desk.

This part of the book was devoted to oscillations and waves, viz. to the theory of oscillations in the broad sense of the word. This theory deals with not only electric oscillations, but also various other oscillations governed by the same laws. But after our remarks made about electric oscillations, it is not surprising that it was radio engineering that provided the theory with the most extensive and interesting material and most complicated and interesting problems. The intense development of the theory of oscillations in 1915-1945 was primarily aimed at meeting the demands of radio engineering.

Later, the question associated with oscillations (and, in particular, self-excited oscillations) has become important for another rapidly developing branch of engineering, viz. *automation* (automatic control of machines and mechanisms, automatic piloting and navigation, telemetric control of spaceships, and so on). Thus, automation also acquired a significant role in the theory of oscillations.

- ❓ I.1. What must be the free-fall acceleration for the length of a pendulum having a period of 2 s to be 1 m? Does the free-fall acceleration on the Earth attain such a value anywhere?
- I.2. What must be the length of a pendulum having a period of 1 min at the latitude of Moscow ( $g = 9.815 \text{ m/s}^2$ )?
- I.3. One of two identical pendulums is placed at the pole and the other on the equator. How many oscillations does the pendulum at the pole complete during the time over which the pendulum on the equator performs 1000 oscillations?
- I.4. The period of the pendulum used for demonstrating the Foucault experiment at the St. Isaac Cathedral in Leningrad is 20 s. The free-fall acceleration  $g$  is  $9.82 \text{ m/s}^2$  at the latitude of Leningrad. Determine the length of the pendulum.
- I.5. A steel ball is released at an altitude  $h$  above a horizontal steel slab. If we neglect the air resistance and energy losses during the impacts against the slab, the ball will periodically jump to the height  $h$  and fall to the slab again. What must be the length  $l$  of a simple pendulum with the same period as the period of motion of the ball?
- I.6. Two balls roll without initial velocity along two grooves from a height  $h$ . One of the grooves is bent in the vertical plane to form an arc of a circle of radius  $R$ , while the other is straight and coincides with the chord of this arc. Assuming that  $h$  is small in comparison

with  $R$ , and neglecting friction, find the periods of time required for the balls to reach the lowermost point (where the bent groove is horizontal). Do these periods of time depend on height  $h$ ? What are the velocities of the balls at the lowermost point?

**I.7.** A pendulum consists of a vessel with water suspended on a long string. Water gradually flows out through the hole at the bottom of the vessel. How will the period of the pendulum change with time? The mass of the vessel should be neglected.

**I.8.** A load stretches a spring by 2 mm. What will be the period of such a simple oscillator?

**I.9.** Why is it impossible to swing a pendulum by pushing it in the same direction twice a period?

**I.10.** At large amplitudes, a pendulum does not manifest isochronism, i.e. its period depends on the amplitude. Does the period increase or decrease with increasing amplitude?

**I.11.** Give an example of self-excited oscillations in a human organism.

**I.12.** Two harmonic voltages of the same frequency and amplitude are applied to the deflection plates of a cathode-ray oscillograph. What will be the motion of the bright spot on the screen of the oscillograph if (a) the voltages are in phase; (b) the voltages are in antiphase; (c) the voltage across the horizontal plates lags behind the voltage across the vertical plates in phase by  $90^\circ$ ?

**I.13.** What is the time required for sound to travel from Moscow to Leningrad (the distance of about 650 km) and for light to propagate from the Moon to the Earth (the distance of about 385 000 km)?

**I.14.** How many times can radio waves run around the Earth along the equator during 1 s?

**I.15.** For measuring huge astronomical distances, *light years* and *parsecs* are used. A light year is the distance traversed by a light wave during a year (365 days). A parsec (abbreviation from parallax-second) is the distance at which the radius of the Earth's orbit (150 000 000 km) is seen at an angle of  $1''$ . Express a light year in terms of parsecs and kilometres.

**I.16.** Why do we hear a loud, deafening thunder when lightning strikes at a close range and a rambling noise in the case of a remote lightning?

**I.17.** Assuming that the lowest frequency perceived by the human ear is 16 Hz, determine the wavelength in water corresponding to it.

**I.18.** A siren having 12 holes in its disc rotates at 700 rpm. Determine the period of acoustic vibrations, their fundamental frequency and the wavelength in air corresponding to it.

**I.19.** A sound track on a record-player starts at a distance of 14 cm from the rotational axis and terminates at a distance of 5.6 cm. The record rotates at a velocity of 32 rpm. What is the length of the period of undulations at the beginning and at the end of the track for a tone having a frequency of 440 Hz?

**I.20.** Why does a toy "telephone" consisting of two membranes connected with a stretched string or wire (Fig. 150) make it possible to communicate through a low voice or even whisper at a distance of a few tens of metres?

**I.21.** How will the form of the interference pattern produced by two coherent sources vibrating in phase change depending on the distance between them? Make an experiment in a water bath by varying the distance between vibrating points.

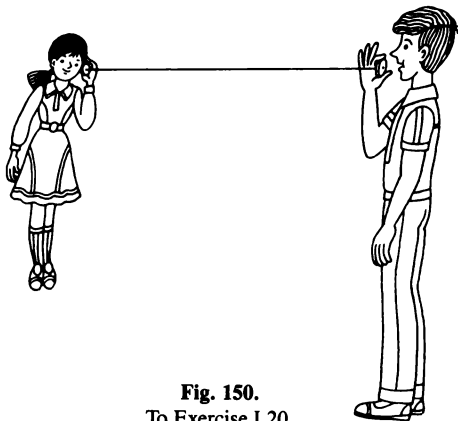
**I.22.** How will the tone of wind instruments change with rising temperature? Is this change the same for metal and wooden trumpets?

**I.23.** What must be the length of a resonance box for a tuning fork having a frequency of 300 Hz?

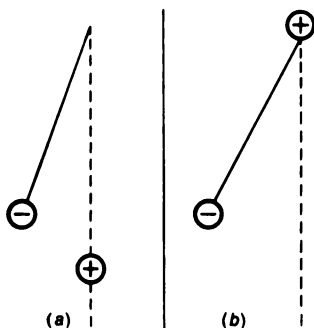
**I.24.** One can whistle by blowing above the opening of a door key. How can the frequency of fundamental tone be determined?

**I.25.** Find the ratio of frequencies corresponding to the fourth overtones for two identical tubes of the same length, one of which is opened and the other is closed at one end.

**I.26.** Why do two prongs of a tuning fork stop vibrating when one of them is touched?



**Fig. 150.**  
To Exercise I.20.



**Fig. 151.**  
To Exercise I.32.

**I.27.** What will be the change in the frequency of a tuning fork if two pieces of wax are attached to the ends of its prongs?

**I.28.** A steel string having a length of 50 cm and a thickness of 0.1 mm is stretched by the load whose mass is 9.68 kg. Find the fundamental frequency and the corresponding wavelength in air. The density of steel is  $7.8 \times 10^3 \text{ kg/m}^3$ .

**I.29.** A steel and a platinum strings of the same cross-sectional area are stretched by identical loads and sound in unison. What is the ratio of their lengths?

**I.30.** Why are bass strings of a piano made in the form of a central steel wire with a wire spiral densely wound on it?

**I.31.** Under what condition can a pier or a damb protect the shore from a storm in the open sea?

**I.32.** A pendulum is made in the form of an ebonite or glass ball suspended on a silk thread. The ball bears a negative charge. How will the period of the pendulum change if the other positively charged ball is brought from below (Fig. 151a) or placed at the point of suspension (Fig. 151b)?

**I.33.** The oscillatory circuit of a radio receiver consists of an induction coil  $L = 5 \times 10^{-8} \text{ H}$  and a capacitor of capacitance  $C$ . For what value of  $C$  will the circuit be tuned to receive radio waves having the wavelength  $\lambda = 94 \text{ m}$ ?

**I.34.** What is the length of a half-wave oscillator whose fundamental frequency is equal to the natural frequency of an oscillatory circuit having a capacitance of  $10^{-10} \text{ F}$  and an inductance of  $10^{-6} \text{ H}$ ?

**I.35.** The inductance of the LC-circuit of a radio receiver is  $2 \times 10^{-5} \text{ H}$ . In what limits should the capacitance vary for tuning the circuit to waves from 35 to 45 m?

## Part Two

# Geometrical Optics

### Chapter 7

## Light Phenomena: General

#### 7.1. Effects of Light

The sensitivity of our organs of vision to light is very high. According to latest measurements, for perceiving light it is sufficient that about  $10^{-17}$  J of radiant energy fall on the eye per second under favourable conditions. In other words, the power sufficient for noticeable light irritation is equal to  $10^{-17}$  W.

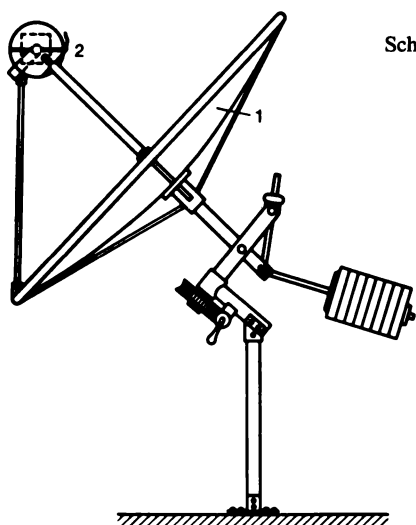
The human eye belongs to the most sensitive instruments capable of registering the presence of light. The action of light on the eye is reduced to a certain chemical process emerging in the sensitive layer of the eye and causing the irritation of the optic nerve and relevant centres in the cerebrum. The chemical effect of light similar to the action on sensitive elements of the human eye can be observed in fading of various dyes in light ("fading of fabrics"). Chemical transformations are observed during the absorption of light by a comparatively small number of light-sensitive materials. Light is absorbed, however, by any body to a certain extent, which can be detected from heating of a body.

The heating of bodies upon the absorption of light is a general process which can easily be realized and used for detection and measurement of radiant energy. The heating by solar radiation is a simple example of such a process. In southern regions where there is a large number of sunny days, the heat obtained as a result of absorption of solar energy can be used for actuating industrial apparatus.

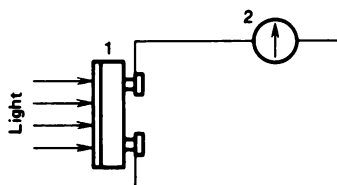
The energy supplied by solar radiation in southern regions during sunny days amounts to more than 1000 J/s per square metre of the surface so that a flat iron tank mounted on the roof of a building can supply hot water to inhabitants during the entire summer season. Concentrating solar rays with the help of a large mirror 1 (Fig. 152) on the surface of receiver 2, it is possible to heat it to a high temperature.

The effect of light can be observed in some electrical phenomena. It was mentioned in Sec. 1.9 of Vol. 2 that the illumination of a metal surface may cause the stripping of electrons from it (*photoelectric effect*). Using special appliances, we can easily detect the electric current emerging under the effect of light. Figure 153 shows the schematic diagram of one such device known





**Fig. 152.**  
Schematic diagram of a solar heat engine  
(cell): 1 — mirror, 2 — receiver.



**Fig. 153.**  
Circuit diagram with a photocell: 1 — pho-  
tocell, 2 — galvanometer.

as a *photoelectric cell* (or photocell). If the roof of a small house could be coated with a substance used in photocells, an electric current having a power of a few kilowatts could be obtained on a bright sunny day at the expense of light energy.

Finally, it should be noted that a direct mechanical action of light can also be observed. It is manifested in the light pressure on the surface of a body reflecting or absorbing light. Making such a body in the form of a movable wing, it is possible to observe the rotation of such a wing under the action of incident light. This remarkable experiment was carried out for the first time in 1900 by P.N. Lebedev in Moscow. Calculations show that the solar radiation incident on a mirror surface of square metre on a bright day acts on it with a force of about  $4 \mu\text{H}$ .

At the present time, new sources of coherent high-intensity radiation have been developed. These are lasers which permit to obtain the light pressure of  $10^6$  atm by concentrating energy to a small surface (see Sec. 22.12). Thus, light may produce various effects. They all confirm that optical radiation carries an energy. The conversion of this energy is observed in all phenomena described above.

The above examples show that effects of light are diverse. However, the role of light as a direct source of energy is comparatively small. Engines based on heating under the action of light are not widely used, and those based on the photoelectric effect still await their development. Experiments show, however, that photocells constructed by using semiconductor germanium and silicon will convert up to 15% of radiant energy incident on them directly into the energy of electric current (solar batteries).

As a matter of fact, the entire energy utilized on the Earth has the light energy (solar radiant energy) as a primary source, but its utilization involves complex energy transformations in fuel accumulated in plants under the action of solar radiation and burnt in heat engines, in water and wind engines, and so on. In most applications of light, of primary importance is not the amount of energy carried by it but its specific properties. In order to clarify the nature of optical phenomena, we must again turn to experiments.

## **7.2. Interference of Light. Colours of Thin Films**

Many of us have been fascinated by the generous splash of colours on the surface of puddles in spring. It can be noticed that similar coloured fringes are observed on the water surface in rivers near ships when oil spots appear. In all these phenomena, the intricate pattern of coloured fringes and especially the play of colours, i.e. the change of colours when we turn the head, is intriguing. This phenomenon is similar to the play of colours on soap bubbles, and is actually identical to it in its physical nature. It can easily be observed in a classroom if we add a drop of kerosene or turpentine oil on the surface of water in a cuvette illuminated by a projection lantern.

The variety of colours in the above patterns is clearly connected with the fact that observations were carried out in white light. Arranging a coloured glass on the path of light, we make sure that instead of coloured fringes, the fringes of the same colour will be observed. These fringes have different intensities and are separated by dark gaps. The shape and arrangement of fringes will not change. If, for example, we use a green glass, the fringes that were green in white light will practically remain unchanged, while red fringes will appear black. The phenomenon will become more clear if we use as a source of monochromatic light the flame of a burner with a piece of asbestos impregnated by common salt solution. Such a flame has yellow colour due to the emission from sodium vapour constituting the salt. The colour is almost uniform. The pattern observed in such a light is formed by bright yellow fringes gradually transformed into deep black fringes. Thus, the pattern consists of alternating light fringes, sending much light to the observer's eye (maxima), and dark fringes directing no light to the observer (minima).

In the experiments described above, we deal with phenomena similar to those described in Secs. 5.2.-5.4 and associated with the interference of waves. We specified the conditions (see Sec. 5.3) under which the superposition of two waves leads to the energy redistribution, i.e. the formation of maxima and minima of energy. In the above optical experiments, we also observed energy redistribution as a result of which the dark regions (minima) and bright regions (maxima) are formed instead of uniform illuminance. In other words, optical experiments revealed the ability of light to produce interference patterns, i.e. the wave nature of light. The fact that maxima for

different colours are located in different regions indicates that different colours correspond to different wavelengths (see Sec. 5.3). Henceforth, we shall consider the interference phenomena in optics in greater detail and use them for the exact determination of optical wavelengths. At the moment, we shall just mention that these wavelengths are shorter than a micrometre.

### 7.3. Brief Information from the History of Optics

The answer to the question about the nature of light waves was obtained on the basis of a long series of observations of peculiarities of optical phenomena. In this case, as it normally happens with the evolution of scientific views, the ideas about the nature of light were modified as new information and data were accumulated.

The ideas about the wave nature of light were developed as far back as in the 17th century by Ch. Huygens and were supported in the 18th century by L. Euler, M.V. Lomonosov and B. Franklin. However, during this period the corpuscular concepts of light, according to which light was identified with a flow of fast particles (I. Newton), remained more substantiated. Only at the beginning of the 19th century, the wave nature of light was reliably substantiated in the works by O. Fresnel and T. Young (see Chaps. 13 and 14). These waves were assumed to be similar to a certain extent to elastic waves responsible for acoustic phenomena. However, two important features distinguish light waves from acoustic waves.

Firstly, light propagates through the space from which air or some other medium has been removed, while sound cannot propagate in vacuum (see Sec. 4.1). The propagation of light in vacuum can be observed in incandescent lamps with air pumped out from the bulbs.<sup>1</sup> Another proof of the ability of light to propagate in vacuum is the possibility to observe the light from the Sun and stars separated from us by enormous space containing less substance per unit volume than the most perfect vacuum devices.

According to modern information, the interstellar space contains about one atom in  $1 \text{ cm}^3$ , while the most thoroughly evacuated devices contain not less than  $10^8$  atoms or molecules in  $1 \text{ cm}^3$ .

Secondly, a distinctive feature of light waves in comparison with sound waves is their huge velocity of propagation. Astronomic observations of eclipses of Jupiter's satellites made by Römer (see Sec. 18.2) showed that the velocity of propagation of light in free space is close to  $300\,000 \text{ km/s}$  ( $3 \times 10^8 \text{ m/s}$ ). The velocity of light in air is practically the same, while sound propagates in air at a velocity equal to one millionth of this value.

---

<sup>1</sup> In most of modern incandescent lamps, the bulb is first thoroughly evacuated, and then filled with a chemically inert gas like nitrogen. However, this is done only to reduce the disintegration of the filament, i.e. to prolong the service life of the lamp. But the light from the filament propagates in the lamps with a perfect evacuation as well.

The tremendous velocity of propagation of light distinguished optical phenomena from all others known in the first quarter of the 19th century. About fifty years later, J. Maxwell established, proceeding from purely theoretical considerations, that any electromagnetic perturbation must propagate just at that velocity. Later, Hertz experimentally obtained electromagnetic waves which propagated at a velocity actually equal to the speed of light.

Further investigations, and above all Lebedev's experiments in which he obtained the shortest electromagnetic waves (6 mm), showed that all basic properties of electromagnetic waves are the same as those of light waves. These important facts led to the conclusion that *light waves are electromagnetic waves* differing from the waves normally used in radio engineering in their small length (less than a micrometre) (see Sec. 6.5).

The electromagnetic nature of light waves explains the emission of electrons by illuminated metals, i.e. the so-called photoelectric effect which was mentioned in Sec. 1.9 of Vol. 2 and will be considered in detail in Chap. 21. There exist a number of other phenomena revealing the relation between light and electromagnetic processes. On the basis of all experimental and theoretical results available at present, we can assume the fact that light waves are electromagnetic waves. Luminous bodies (like the Sun) emit electromagnetic (primary) waves. Meeting a body on its way, such a primary wave causes forced oscillations of electrons constituting the body, which become the sources of secondary electromagnetic waves. The variety of optical phenomena, the colours and silhouettes of objects are due to the superpositions of primary and secondary waves. As was mentioned earlier, many features of wave phenomena are similar for wave processes of different nature. Therefore, while studying basic principles and concepts of optics, we shall make use of the information about waves contained in Chaps. 4-6. The experimental results accumulated by the 20th century led to the conclusion that light possesses, along with wave properties, the corpuscular properties as well (quanta of light, or photons, Sec. 21.3). At present, the quantum theory combines the wave and corpuscular ideas about light into a single whole, just like it incorporates the wave and corpuscular conceptions concerning electrons, atoms and other particles (see Sec. 22.17).

## Chapter 8

# Photometry and Lighting Engineering

### 8.1. Radiant Energy. Luminous Flux

As it was noted in Sec. 7.1, various effects of light are primarily due to the existence of radiant (luminous) energy.

We directly perceive light as a result of the action of radiant energy absorbed by the sensitive elements of the eye. The same takes place in any receiver capable to respond to light, like a photoelectric cell, thermocouple or a photographic plate. Therefore the measurement of light is reduced to the measurement of radiant energy or to the measurement of quantities somehow related to it. The branch of optics dealing with methods and techniques of measuring radiant energy is called *photometry*.

Let us mentally isolate a small area element  $\sigma$  on the path of light propagating from a source  $S$  (Fig. 154). A certain radiant energy  $W$  will pass through this area element over time  $t$ . In order to measure this energy, we should assume that this area element has the form of a film covered with a substance (say, soot) absorbing completely the radiant energy incident on it, and measure the energy absorbed from the heating of the film. The ratio

$$\Phi = W/t \quad (8.1.1)$$

indicates the energy passing through the area element per unit time and is known as the *radiation flux* (or *radiating power*) through the area element  $\sigma$ . It should be recalled that the power transported by a light wave through a unit area element is called the *intensity* of wave (see Sec. 4.7).

The radiation flux is measured in ordinary units of power, viz. watts, while the radiant intensity is measured in watts per square metre. The eye plays an exceptionally important role for the perception and utilization of luminous energy. For this reason, along with the energy characteristic of light, the estimate based on direct perception of light by the eye is also used. *The radiation flux estimated according to the visual sensation is termed the luminous (light) flux.*

Thus, in optical measurements two systems of notations and two systems of units are used. One of them is based on the estimation of light from the energy viewpoint, while the other, on the estimation according to the visual sensation.

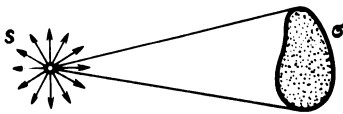
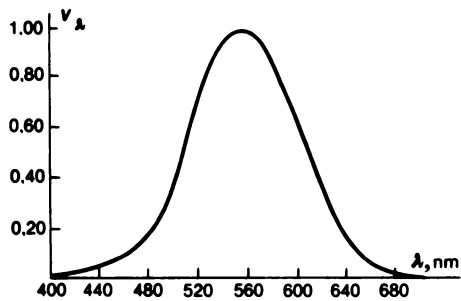


Fig. 154.

Luminous flux emitted by source  $S$  and passing across surface area element  $\sigma$ .

Fig. 155.

Curve describing relative spectral sensitivity of eye.



Since the sensitivities of the eye to light of different wavelengths (different colours) are different, the energy estimate may differ considerably from the estimate of a luminous flux according to visual sensation. For example, for the same radiating power, the visual sensation from green rays will be about 100 times stronger than from red or bluish-violet rays. Therefore, for a visual evaluation of luminous fluxes, we must know the sensitivity of the eye to light of different wavelengths, or the so-called *curve of relative spectral sensitivity of eye*, shown in Fig. 155. This curve shows the relative sensitivity  $v_\lambda$  of the human eye depending on wavelength  $\lambda$ . If the sensitivity of the eye corresponding to wavelength  $\lambda = 555 \text{ nm} = 5550 \text{ \AA}$ <sup>1</sup> (green light) is taken for unity, for longer and shorter waves the sensitivity rapidly decreases as follows from the curve. For example, for  $\lambda = 510 \text{ nm}$  and for  $\lambda = 610 \text{ nm}$ , the sensitivity is equal to 0.5 (i.e. reduced to a half). For  $\lambda = 470 \text{ nm}$  (blue) and  $\lambda = 650 \text{ nm}$  (orange-red) the sensitivity constitutes about 0.1, while for  $\lambda = 430 \text{ nm}$  (bluish-violet) and  $\lambda = 675 \text{ nm}$  (red), it amounts to about 0.01, and so on.

The sensitivity curves for eyes of different persons slightly differ, especially in the region of low sensitivities. The curve in Fig. 155 was obtained on the basis of numerous measurements. It characterizes the sensitivity of an average normal eye and has been ratified by the International Committee on Standards.

## 8.2. Point Sources of Light

All problems associated with determining optical quantities are solved in an especially simple way when a light source emits uniformly in all directions. An example of such a source is an incandescent metal ball. Such a ball emits light uniformly in all directions. The luminous flux is then distributed uniformly in all directions. This means that the action of a

<sup>1</sup> The symbol  $\text{\AA}$  indicates the distance to  $10^{-10} \text{ m} = 0.1 \text{ nm}$ . This unit of length is called the angstrom unit (or angstrom) in honour of the Swedish scientist Knut Johan Angström (1814-1874).

source on a light receiver will depend only on the distance between the receiver and the centre of the luminous ball and will not depend on the direction of the radius drawn from the centre of the ball to the receiver.

In many cases, the action of light is studied at a distance  $R$  which is so much larger than the radius  $r$  of the luminous ball that the size of the latter can be ignored. Then it can be assumed that the light is emitted as if from a single point, viz. the centre of the luminous ball. In such cases, the light source will be referred to as a *point source*.

Naturally, a point source is not a point in the geometrical sense, but rather has finite dimensions like any other physical body. The source of radiation of a vanishingly small size has no physical meaning since such a source should emit an infinitely high power per unit area, which is impossible.

Moreover, a source that can be treated as a point source should not always be small. What is important is the ratio between the size of the source and the distances at which its action is being investigated, and not its absolute size. For example, stars are the best examples of point sources for all practical problems. Although they are huge in size, their distances from the Earth exceed their size many times.

It should also be recalled that a uniformly glowing ball is a prototype of a point source. Therefore, a light source emitting light nonuniformly in different directions is not a point source although it can be very small in comparison with the distance to the point of observation.

Let us define more rigorously what we mean by uniform radiation in all directions. For this we shall need the concept of the solid angle  $\Omega$  which is equal to the ratio of the area  $\sigma$  of the surface cut in a sphere by a cone with the apex at point  $S$  to the squared radius  $r$  of the sphere (Fig. 156);

$$\Omega = \sigma/r^2. \quad (8.2.1)$$

This ratio does not depend on  $r$  since as  $r$  increases, the area  $\sigma$  of the surface cut by the cone increases in proportion to  $r^2$ . If  $r = 1$ ,  $\Omega$  is numerically equal to  $\sigma$ , i.e. the solid angle is measured by the surface cut by the cone in the sphere of unit radius. The unit of solid angle is a steradian<sup>2</sup>

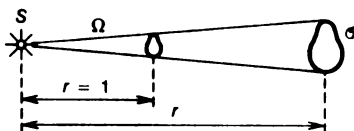


Fig. 156.

Solid angle  $\Omega$  is equal to the ratio of the area  $\sigma$  of the surface cut from a sphere by the cone with the apex at point  $S$  to the squared radius  $r$  of the sphere:  $\Omega = \sigma/r^2$ .

<sup>2</sup> Steradian is a spatial radian. It follows from the text that it is defined in complete analogy with a radian which is a unit of plane angle.

(sr), viz. a solid angle corresponding to the unit surface area on a sphere of unit radius. The solid angle embracing the entire space around a source is equal to  $4\pi$  sr since the total surface area of a sphere having a unit radius is  $4\pi$ .

The total radiation from a source is distributed in a solid angle of  $4\pi$  sr. The radiation is referred to as *uniform*, or *isotropic*, if the same power is emitted within equal solid angles isolated in any direction. Naturally, the smaller the solid angles in which we compare the power emitted by a source, the higher the degree of accuracy with which we verify the uniformity of radiation.

Thus, a *point source* is a source having a size which is small in comparison with the distance to the point of observation and emitting a luminous flux uniformly in all directions.

### 8.3. Luminous Intensity and Illuminance

The total luminous flux characterizes the radiation propagating from a source in all directions. For practical purposes, however, it is often important to know the flux in a certain direction or a flux incident on an area element rather than the total luminous flux. For example, for a car driver, it is important to obtain a powerful beam of light in a comparatively narrow solid angle containing a small region of a highway. For a person sitting at a writing desk, it is important that light should fall on the desk, or even the part of the desk being used by him. Accordingly, two auxiliary concepts have been introduced, viz. the *luminous intensity* ( $I$ ) and *illuminance* ( $E$ ).

The luminous intensity is the luminous flux referred to a solid angle of one steradian, i.e. the ratio of the luminous flux  $\Phi$  contained in a solid angle  $\Omega$  to this angle:

$$I = \frac{\Phi}{\Omega}. \quad (8.3.1)$$

The illuminance is the luminous flux per unit area, i.e. the ratio of the luminous flux  $\Phi$  incident of the surface area  $\sigma$  to this area:

$$E = \frac{\Phi}{\sigma}. \quad (8.3.2)$$

Formulas (8.3.1) and (8.3.2) clearly define the average luminous intensity and the average illuminance. These values will be the closer to the actual values, the more uniform is the flux and the smaller are  $\Omega$  and  $\sigma$ .

Obviously, using a source emitting a certain luminous flux we can obtain various luminous intensities and various illuminances. Indeed, if we direct the entire flux or its larger part into a small solid angle, then a large luminous intensity will be obtained in the direction isolated by this angle. For example, the most part of the flux emitted by the electric arc in a flood-



light is concentrated in a very small solid angle, and a very large luminous intensity is obtained in the corresponding direction. The same effect is attained (to a smaller extent) with the help of headlights of a motor car. If the luminous flux from a source is concentrated on a small area with the help of reflectors or lenses, a considerable illuminance can be attained. This effect is used when a preparation observed through a microscope should be well illuminated. The reflector of a lamp illuminating a working place serves for the same purpose.

According to formula (8.3.1), the luminous flux  $\Phi$  is equal to the product of the luminous intensity  $I$  and the solid angle  $\Omega$  in which it propagates:

$$\Phi = I\Omega.$$

If the solid angle  $\Omega = 0$ , i.e. the rays are strictly parallel, the luminous flux is also equal to zero. This means that a strictly parallel beam of light does not carry any energy, i.e. has no physical meaning. In other words, a strictly parallel beam of light cannot be realized in any experiment. This is a purely geometrical concept. Nevertheless, parallel light beams are widely used in optics. As a matter of fact, small deviations from the parallelism of light rays, which are of principal importance from the point of view of energy, are practically immaterial for the questions connected with the passage of light rays through optical systems. For example, the angles at which the rays from a remote star are incident on an observer's eye or a telescope are so small that they cannot be measured by existing methods. These rays practically do not differ from parallel rays. However, the angles differ from zero, and for this reason we can see stars. Sharply focussed light beams, i.e. rays of light with a very small divergence, have been obtained recently with the help of lasers (see Sec. 22.12). However, even in this case the angle between the light rays has a finite value.

#### 8.4. Laws of Illumination

Formulas (8.3.1) and (8.3.2) show that the quantities  $E$  and  $I$  are connected with each other.

Let a point source  $S$  illuminate a small area element  $\sigma$  at a distance  $R$  from the source (Fig. 157).

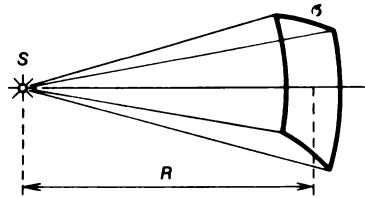
We construct a solid angle  $\Omega$  whose vertex lies at the point  $S$  and which rests on the edges of the area element  $\sigma$ . It is equal to  $\sigma/R^2$ . The flux emitted by the source into this solid angle will be denoted by  $\Phi$ . Then the luminous intensity  $I = \Phi/\Omega = \Phi R^2/\sigma$ , and the illuminance is  $E = \Phi/\sigma$ . Hence,

$$E = I/R^2, \quad (8.4.1)$$

i.e. the illuminance of an area element is equal to the luminous intensity

Fig. 157.

The illuminance of the surface area element  $\sigma$  perpendicular to the axis of the luminous flux is determined by the luminous intensity and the distance  $R$  from point source  $S$  to the area element.



divided by the squared distance from the point source. Comparing the illuminance of the area elements located at different distances  $R_1$  and  $R_2$  from the point source, we find that  $E_1 = I/R_1^2$  and  $E_2 = I/R_2^2$  and so on, or

$$E_1/E_2 = R_2^2/R_1^2. \quad (8.4.2)$$

Therefore, the illuminance is inversely proportional to the squared distance from the area element to the point source. This is the so-called *inverse-square law*.

If the surface element  $\sigma$  were not perpendicular to the axis of the light flow but were arranged at an angle  $\alpha$  to it, its area would be  $\sigma = \sigma_0/\cos \alpha$  (Fig. 158), where  $\sigma_0$  is the surface element subtending the same solid angle perpendicular to the axis of the light beam, so that  $\Omega = \sigma_0/R^2$ . It was assumed that surface elements  $\sigma$  and  $\sigma_0$  are so small and so remote from the source that the distance  $R$  from all points of these elements to the source can be assumed to be the same, and the rays form with the normal to the surface element  $\sigma$  the same angle  $\alpha$  (*angle of incidence*) at all points.

Then the illuminance of the surface element  $\sigma$  is

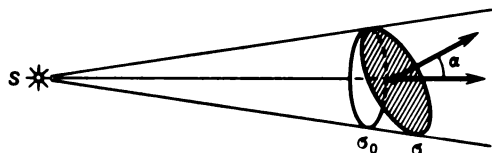
$$E = \frac{\Phi}{\sigma} = \frac{\Phi \cdot \cos \alpha}{\sigma_0} = \frac{\Phi \cos \alpha}{\Omega \cdot R^2} = \frac{I \cos \alpha}{R^2}. \quad (8.4.3)$$

Therefore, the illuminance produced by a point source on a certain surface element is equal to the *luminous intensity multiplied by the cosine of the angle of incidence of light on the surface element and divided by the squared distance to the source*.

The inverse-square law is strictly observed for point sources. If, however, the size of a source is not very small in comparison with the distance to the illuminated surface, relation (8.4.1) does not hold, and the illuminance decreases slower than in proportion to  $1/R^2$ . In particular, if the size of a luminous surface is large in comparison with  $R$ , the illuminance practically remains unchanged upon a variation of  $R$ . The smaller the size  $d$  of

Fig. 158.

The illuminance of the surface element  $\sigma$  is proportional to the cosine of angle  $\alpha$  formed by the normal to the surface with the direction of the luminous flux.



the source in comparison with  $R$ , the more valid the inverse-square law. For instance, for the ratio  $d/R \leq 1/10$ , the illuminance calculated by (8.4.1) is in good agreement with the measured value. Thus, the inverse-square law can be assumed to be practically applicable if the size of a source does not exceed 0.1 of the distance to the illuminated surface.

Formula (8.4.3) shows that illuminance also depends on the angle at which light rays are incident on this surface.

### 8.5. Units of Photometric Quantities

The base unit of the photometric system is the unit of luminous intensity. This unit is conditional: the luminous intensity of a certain standard source was taken for the unit of luminous intensity. At first it was assumed that the source having a luminous intensity  $I = 1$  is the flame of a candle manufactured by a strictly standard method. However, such a standard source proved to be inconvenient since its luminous intensity somewhat changes as the snuff is formed and, besides, it depends on the temperature and humidity of air. Many other sources were proposed as standards of luminous intensity, including standard incandescent electric lamps whose samples are stored at large-scale state measuring laboratories and controlled by comparing with other samples of this type.

The unit of luminous intensity is a *candela* (cd). A candela is the luminous intensity in a given direction from a source emitting a frequency of  $540 \times 10^{12}$  Hz (the wavelength in vacuum is 555 nm) and having a radiant intensity of 1/683 W/sr. Candela is a base unit of the International System of Units (SI).

Standards in the form of incandescent electric lamps are not sufficiently permanent and cannot be reproduced with a high degree of accuracy in the case of their breakdown. For this reason, a new standard has been introduced by international convention, which can be reproduced accurately. It is a special vessel in which chemically pure platinum is fused. A narrow refractory tube heated to the temperature of platinum is inserted into it. Light is emitted by the inner surface of the tube through its open end. The temperature of the crystallizing pure platinum has a strictly definite value equal to 2042 K. The luminous intensity of the light emitted at this temperature from the surface of the tube (whose area is  $(1/60)\pi \text{ cm}^2$ ) in the direction of its axis is strictly constant. *This luminous intensity is equal to a candela.*

The unit of luminous flux is a *lumen* (lm). A lumen is the luminous flux emitted by a point source whose luminous intensity is one candela within a unit solid angle (i.e. an angle of 1 sr). For the radiation corresponding to the maximum spectral sensitivity of the human eye ( $\lambda = 555 \text{ nm}$ ), the luminous flux is equal to 683 lm if its radiant intensity is 1 W/sr.

The unit of illuminance is defined as the illuminance of a  $1\text{-m}^2$  surface over which a luminous flux of 1 lumen is uniformly distributed. This unit of illuminance is called a *lux* (lx). The illuminance of one lux is obtained

on the surface of a sphere having a radius of 1 m with a light source of 1 cd placed at its centre. The values of illuminance for some typical cases are given in Table 1.

**Table 1.** Illuminance (in Luxes) in Some Typical Cases

| Illuminance                                     | Illumination |
|-------------------------------------------------|--------------|
| By direct solar rays at noon (mean latitudes)   | 100 000      |
| During shooting in a film studio                | 10 000       |
| In an open place on cloudy day                  | 1 000        |
| In a well-lit room not very far from the window | 100          |
| On a working table for fine workmanship         | 100-200      |
| Required for reading                            | 30-50        |
| On a cinema screen                              | 20-80        |
| From a full Moon                                | 0.2          |
| From a night sky at moonless night              | 0.0003       |

With the discovery of lasers which have a very high radiant intensity, it has become possible to produce much larger illuminance although during short intervals of time. The property of laser to emit light beams of a small divergence is also very important. Owing to this property, the entire laser radiation can be concentrated to a spot having an area of about  $10^{-8} \text{ cm}^2$ . A small laser with an energy of 0.1 J emitted per pulse lasting for  $10^{-8} \text{ s}$  creates within such spot an “enormous” density of a power equal to  $10^{13} \text{ W/cm}^2$  or  $10 \text{ TW/cm}^2$  during a pulse.<sup>3</sup> It should be noted that the power of all electric power stations on the globe amounts to about 1 TW. It can easily be calculated that the illuminance created by such a laser within a small spot is about  $10^{20} \text{ lx}$  for light with a wavelength  $\lambda = 555 \text{ nm}$ , i.e. is  $10^{15}$  times higher than the maximum illuminance produced by the Sun.

### 8.6. Brightness of Sources

Till now, we have considered only point sources of light. In actual practice, the light sources are normally extended, i.e. while observing them from a given distance, we can distinguish their shape and size. In order to characterize extended sources even in the simple case when they are in the form of uniformly glowing balls, a single quantity, viz. their luminous intensity, is insufficient. Indeed, let us imagine two luminous balls emitting light uniformly in all directions and having the same luminous intensity but different diameters. The illuminance produced by each such ball at the same distance from their centres will be the same. However, these balls will seem to be quite different light sources: the smaller ball appears to be brighter

<sup>3</sup> The prefix “tera” (T) is derived from the Greek *teras* meaning a monster.

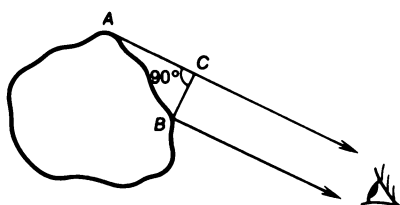


Fig. 159.

Relation between the area of the actual emitting surface ( $AB$ ) and that of the surface visible in a given direction ( $BC$ ).

than the larger one. This is due to the fact that for the same luminous intensity, the radiating surface of one ball is larger than that of the other ball, and hence the luminous intensity emitted per unit surface area of the source will be different in the two cases. It should be noted that when we consider a light source, we are interested in the size of the visible surface rather than in the area of the emitting surface. In other words, the projection of the emitting surface on the plane perpendicular to the direction of observation is important for us (Fig. 159).

Thus, we arrive at the conclusion that in order to characterize the properties of an extended light source, we must know the *luminous intensity per unit area of the visible surface of the source*. This photometric quantity is known as the *brightness* of the source. We shall denote it by  $L$ . If the luminous intensity of a source is  $I$  and the area of the visible surface is  $\sigma$ , the brightness of this source is

$$L = \frac{I}{\sigma}. \quad (8.6.1)$$

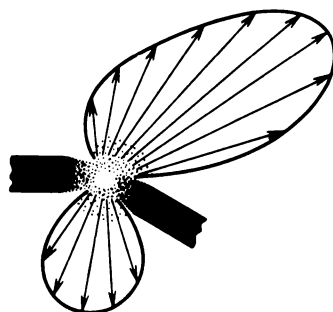
Using formula (8.3.1), we can also write

$$L = \frac{\Phi}{\Omega \sigma}, \quad (8.6.2)$$

i.e. we can assume that the *brightness of a source is equal to the luminous flux emitted per unit area of the visible surface of the source in a unit solid angle*.

The brightness of some regions of a light source may differ from the brightness of other regions. For example, different parts of the flame of a candle, lamp, etc. are characterized by sharply differing brightnesses. Besides, the brightness depends on the direction in which the radiation from a source is analyzed. This is due to the fact the luminous intensity of many sources depends on direction. An electric arc, for example, does not emit light in some directions at all (Fig. 160).

Thus, the brightness can serve as a characteristic of radiation from a region of the surface of a source in a given direction. At the same time, brightness is very important since, as will be shown below, it is a photometric quantity to which the eye responds directly.

**Fig. 160.**

Brightness of an electric arc, which is proportional to the length of the arrows in the figure, depends on the direction of radiation.

The unit of brightness is a *candela per square metre*. This is the brightness of a luminous surface producing a luminous intensity of 1 cd from each square metre of its surface in a direction perpendicular to this surface.

The characteristics of brightness of various luminous bodies are given in Table 2.

**Table 2.** Brightness of Some Light Sources (in  $\text{cd/m}^2$ )

| Light source                                    | Brightness                        |
|-------------------------------------------------|-----------------------------------|
| The Sun                                         | $1.5 \times 10^9$                 |
| A capillary of a superhigh-pressure mercury arc | $(1.2-1.5) \times 10^9$           |
| The crater of a carbon arc                      | $1.5 \times 10^8$                 |
| The metal filament of an incandescent lamp      | $1.5 \times 10^8 - 2 \times 10^6$ |
| The flame of a kerosene lamp                    | $1.5 \times 10^4$                 |
| The flame of a steaming candle                  | $0.5 \times 10^4$                 |
| Moonless night sky                              | $10^{-4}$                         |
| The smallest brightness perceived by the eye    | $10^{-6}$                         |

Sources of light with a large brightness (above  $1.6 \times 10^5 \text{ cd/m}^2$ ) cause a sensation of pain in the eyes. Various appliances are used to protect human eyes from the effect of very bright sources. For example, it is harmful and even painful for the eye to watch the red-hot filament of an incandescent lamp. If, however, the bulb of a lamp is made of frosted or opal glass, or covered by a shade in the form of opaque sphere, the luminous flux emitted by it emanates from a larger area. As a result, the brightness decreases, while the luminous flux practically remains unchanged, and hence the illuminance produced by the lamp does not change either.

### 8.7. Problems of Lighting Engineering

Having introduced the basic photometric quantities characterizing light sources and surfaces being illuminated, we can consider one of the most important practical problems, viz. the evaluation of optimal illumination of residential and industrial buildings as well as the places of social activity.

The branch of physics and engineering dealing with this problem is known as *lighting engineering*. It is devoted to optimal utilization of day-

light in rooms and halls, which can be attained by the properly designed size and the rational arrangement of windows. Another important and complicated problem of lighting engineering is the design of the sources of artificial light, which produce the required illumination at the minimum cost and energy expenditures. In large countries like the USSR, where the energy consumption for illumination purposes are very large, the problems of rational illumination are very important for national economy.

An appropriately designed illumination ensures normal unstrained functioning of the eyes. Therefore, labour productivity increases and the quality of production is improved by using a favourable illumination. Thereby, the normal vision of workers is preserved, hygienic working conditions are created, and the number of accidents is reduced.

Various types of lighting equipment are used for illumination. They consist of a light source (lamp) and lighting fittings. Lighting systems of various kinds cannot increase the total luminous flux which characterizes the emitting source. They play, however, an important role in the redistribution of luminous fluxes and their concentration in the required direction. In this way, the luminous intensity can be increased in a required direction and, accordingly, reduced in other directions.

Another important problem often encountered in lighting engineering is the creation of a uniform illumination of large areas.

Below we shall consider various methods used for solving these problems.

### 8.8. Appliances for Concentrating Luminous Flux

A luminous flux can be effectively concentrated in a given direction with the help of mirrors of a definite shape, used in floodlights, viz. lighting devices intended for the illumination of remote objects. The surface of the mirrors used for this purpose normally have a parabola in any longitudinal section (Fig. 161). Line  $AB$  is known as the parabola axis, and point  $F$  is its focus. The surface is called a *paraboloid*. The axis common to all parabolic sections is the *paraboloid axis* and  $F$  is its *focus*. The geometrical

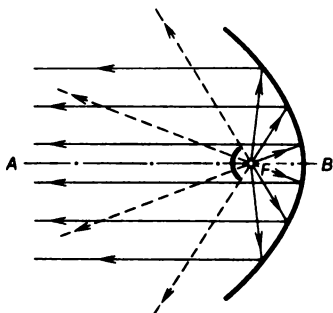


Fig. 161.

Longitudinal section of a floodlight mirror.

properties of a paraboloid are such that a ray emerging from the focus and reflected at any point of the paraboloid becomes parallel to the paraboloid axis. If we place a point light source at the focus of a paraboloid, we obtain a parallel beam of light with the cross section equal to the collimating aperture of the mirror. But since any point source has finite (although very small) dimensions, we can obtain a light beam only more or less close to a parallel beam.

For a ray emerging from a floodlight to have a small divergence angle, i.e. be as close as possible to a parallel beam, the source of light located at the focus of the floodlight must be as small as possible. Clearly, this light source should be bright.

The divergence of the beam of light obtained from an arc lamp and a parabolic mirror having a diameter of about 2 m is about  $1^\circ$ . Optical quantum oscillators, or lasers, produce a much narrower light beams. Using lasers with the beam cross-sectional area of the order of  $1 \text{ cm}^2$ , it is possible to obtain light beams with a divergence of the order of a few angular minutes. With the help of such beams, optical location of the Moon was realized: a region on the Moon surface was illuminated with the help of a laser with such a brightness that the reflected beam could be registered by a sensitive receiver of radiation.

### 8.9. Reflectors and Scatterers

Besides the problems involving the concentration of luminous flux, the need to distribute this flux over a large area for creating a uniform and moderate illuminance often arises. For this purpose, the luminous flow is made to be reflected and scattered by some surfaces. However, we must take into account the fact that only a part of a flow is reflected by or transmitted through a body, while the rest of the light is inevitably absorbed.

The fact that we can see bodies is due to different reflective, refractive and absorbing abilities of these bodies. If a body reflects light stronger than the bodies surrounding it, it appears as a bright body against a dark background. On the other hand, if a body reflects less light than the surrounding bodies, it will appear as dark. For example, white paper reflects more light than grey cardboard, and a piece of cardboard on a sheet of paper appears dark for us. If, however, we put the same piece of cardboard on black velvet (whose reflectivity is very low), it appears bright. A body reflecting light to the same extent as its background cannot be distinguished from it.

Transparent bodies are seen partially in the reflected light and partially in the light transmitted by them. For example, examining such an apparently simple body as a cut glass stopper of a carafe, we deal with a number of



complex phenomena: the light is partially reflected from the faces of the stopper (or scattered if the faces are frosted) and a part of the light passes through the stopper, having been refracted at its surface. If a completely transparent body is immersed in a liquid with the same refractive index as for the given body, the latter will become invisible since the light rays will pass through it without changing their direction or intensity.

The absorption of light leads to losses in a light flow whose energy is spent for heating a body which has absorbed it. As a rule, the absorption of light should be avoided. However, sometimes it is necessary to create a dark background or eliminate light in undesired direction. For this purpose strongly absorbing coatings (like the blackening of surfaces in optical instruments) are used. Absorption is characterized by the *absorption coefficient*  $\alpha$  which is equal to the ratio of the luminous flux  $\Phi_\alpha$  absorbed by a body to the luminous flux  $\Phi_i$  incident on it:

$$\alpha = \Phi_\alpha / \Phi_i. \quad (8.9.1)$$

The reflection of a luminous flux is characterized by the *reflection coefficient*  $\rho$  equal to the ratio of the reflected flux to the incident flux  $\Phi_i$ :

$$\rho = \Phi_\rho / \Phi_i. \quad (8.9.2)$$

Finally, the transmission of light is characterized by the *transmission coefficient*  $\tau$  equal to the ratio of the luminous flux transmitted by a body to the incident flux  $\Phi_i$ :

$$\tau = \Phi_\tau / \Phi_i. \quad (8.9.3)$$

According to the energy conservation law, we have

$$\Phi_i = \Phi_\alpha + \Phi_\rho + \Phi_\tau$$

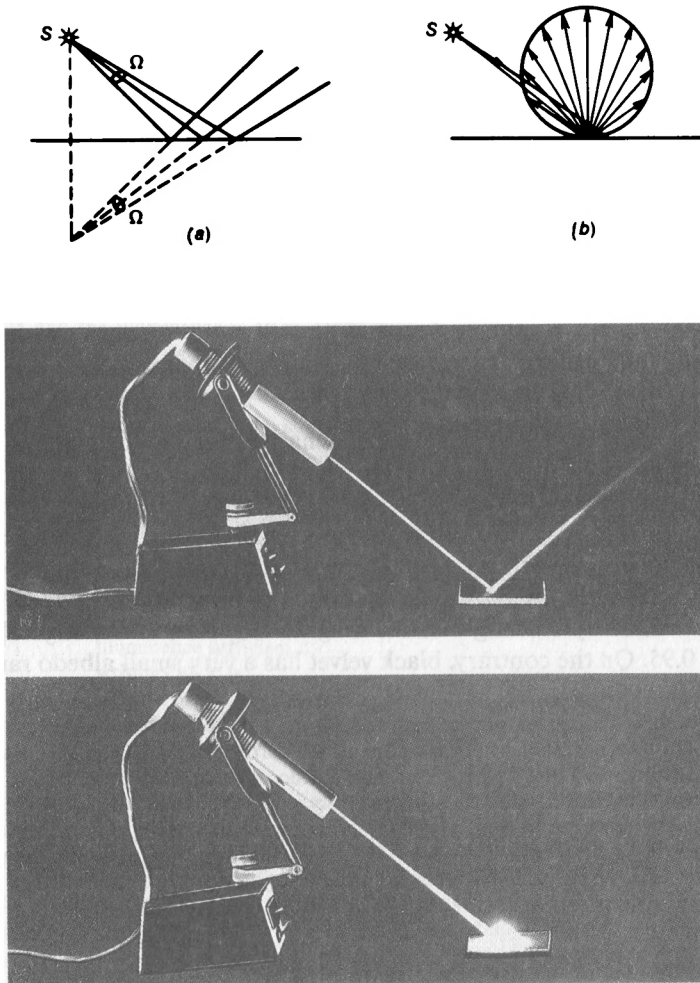
whence it follows from (8.9.1)-(8.9.3) that

$$\alpha + \rho + \tau = 1, \quad (8.9.4)$$

Thus, the *sum of the coefficients of absorption, reflection and transmission is equal to unity*. The coefficients  $\alpha$ ,  $\rho$  and  $\tau$  normally depend on the colour (wavelength) of light.

Both in reflection and transmission of a luminous flux, *direct* and *diffuse* (scattered) reflection and transmission should be distinguished.

For a mirror reflection from a plane surface, the solid angle of a luminous flux does not change (Fig. 162*a*, *c*). With a diffuse reflection, the solid angle within which the light propagates increases (Fig 162*b*, *d*). This increase can be significant to a certain extent depending on the properties of the scattering surface. Similarly, direct transmission is characterized by an invariable solid angle for a luminous flux passing through the body, like for the transmission of light through a plane-parallel plate (Fig. 163*a*).


**Fig. 162.**

Reflection of luminous flux from a plane surface: (a) directional reflection, (b) diffuse reflection, diagram (b) does not change with the angle of incidence of the primary beam, (c) directional (mirror) reflection, a parallel beam of light incident on a polished metal surface generates a clear-cut reflected ray, (d) diffuse reflection, a parallel beam of rays incident on a white sheet of paper is reflected in all directions.

On the contrary, a diffuse transmission is accompanied by a more or less considerable increase of the solid angle of the luminous flux. An example of a diffusely reflecting surface is matt paper, while an example of a diffusely transmitting material is frosted (opal) glass. Frosted glass is simultaneously a diffuse reflector and a diffusely transmitting medium.

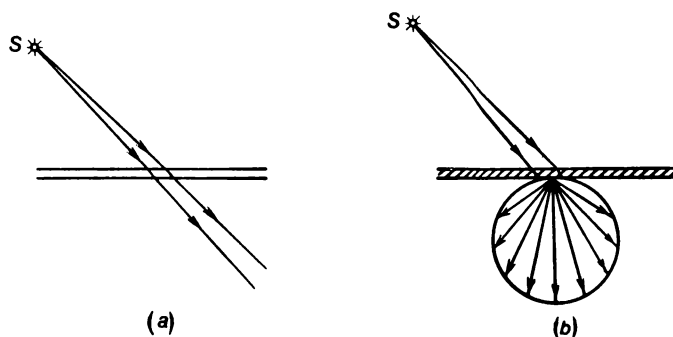


Fig. 163.

Transmission of light through a plane-parallel plate: (a) directional transmission, (b) diffuse transmission. The latter diagram does not change with the angle of incidence of the primary beam.

Scattering properties of a surface are characterized by diagrams similar to those shown in Figs. 162*b* and 163*b*, where the lengths of arrows indicate a fraction of luminous flux scattered in a certain direction. Diffusely reflecting surfaces may also differ in the reflection coefficients which are known as *albedo* for such surfaces. For example, the albedo of white drawing paper is about 0.70-0.80. Surfaces covered by magnesium oxide (a white powder obtained by burning metallic magnesium) have a very high albedo of about 0.95. On the contrary, black velvet has a very small albedo ranging from 0.01 to 0.002.

The albedo of the Earth's surface and its dependence on the colour (wavelength) is of great importance for observations of the surface of the Earth from aeroplanes, and especially for aerial photography. Various types of soil have albedos ranging from 0.2 to 0.4, the higher values corresponding to the region of orange-red colour. Sands are poor reflectors (they reflect about 0.1) in the violet region (which is important for aerial photography), but their albedo increases to 0.5 in the red region. The albedo of grass and leaves is up to 0.50 in the yellow-green region (especially in autumn). Snow also has a very high albedo reaching 0.85 for all wavelengths.

### 8.10. Brightness of Illuminated Surfaces .

Cinema screens, painted ceilings and walls, decorations etc. are diffusely reflecting surfaces.

Being illuminated, such surfaces play the role of extended light sources with large surfaces and normally of moderate brightness. In this respect, they successfully compliment to small glowing sources of light (incandescent lamps, gas-discharge lamps, candles, etc.) which have small surface areas and high brightness.

The brightness of such an illuminated surface is obviously *proportional to its illuminance*. Indeed, the higher the illuminance, i.e. the larger luminous flux is incident per unit area of the surface, the larger the flux reflected by this surface, and hence the brightness of the illuminated surface.

The brightness of an illuminated surface is, besides, the larger, the higher its albedo, i.e. the larger the fraction of luminous flux incident on it, which is scattered by the surface. Thus, the brightness  $L$  of an illuminated surface should be proportional to the product of the illuminance  $E$  and the albedo  $\rho$ , i.e.  $L \propto \rho \cdot E$ . Depending on the scattering diagram, brightness may be different for different directions, and its calculation is a complicated problem. The problem is simplified if a surface scatters light uniformly in all directions. In this case, the brightness is the same in all directions and is given by

$$L = \rho \cdot E / \pi. \quad (8.10.1)$$

If the illuminance  $E$  is expressed in luxes, the obtained brightness will be in candelas per square metre.

Let us determine, for example, the brightness of a cinema screen if its reflection coefficient  $\rho = 0.75$  and the illuminance is 50 lx. Using formula (8.10.1), we obtain

$$L = \frac{0.75 \times 50}{\pi} \approx 12 \text{ cd/m}^2.$$

The brightness of some illuminated surfaces frequently encountered in everyday life are given in Table 3.

**Table 3.** Brightness of Some Illuminated Surfaces (in  $\text{cd/m}^2$ )

| Surface                                                                    | Brightness    |
|----------------------------------------------------------------------------|---------------|
| Of a cinema screen                                                         | from 5 to 20  |
| Of a white sheet of paper at illuminance sufficient for writing (30-50 lx) | from 10 to 15 |
| Of a snow under direct solar rays                                          | 3000          |
| Of the Moon                                                                | 2500          |

### 8.11. Photometry and Measuring Instruments

Photometric quantities can be measured directly with the help of the eye (visual methods) or with the help of a photoelectric cell or a thermopile (objective methods). The instruments for measuring photometric quantities are called *photometers*.

Visual methods are based on the property of the human eye to establish reliably the equality of the brightness of two adjacent surfaces. At the same time, it is difficult to estimate by eye the ratio of the brightness of two surfaces. For this reason, in all visual photometers the role of the eye is reduced to the establishment of the equality of brightnesses of two adjacent surface elements illuminated by the sources being compared.

Since the compared surfaces are made diffusely reflecting, the equality of their brightness indicates, according to what was said in the previous section, that their illuminances are equal. The illuminance of the surface element on which the light from a more powerful source is incident is weakened by a certain way in a known proportion. Having established the equality of the brightness of the two elements and knowing the factor by which the light of one of the sources has been weakened, the luminous

intensities of the two sources can be compared quantitatively. Therefore, each photometer must have two adjacent optical fields one of which is illuminated by only one source and the other, by the other source. The shapes of the fields being compared may be different. In most cases, they are made in the form of two adjacent semicircles (Fig. 164a) or two concentric circles (Fig. 164b). The two fields being compared must be illuminated by their own sources at the same angle and the observer's eye should examine both fields at the same angle of vision.

Figure 165 shows the schematic diagram of a simple photometer. The light from sources  $S_1$  and  $S_2$  to be compared illuminates the white faces of prism  $ABC$  placed in a blackened tube. The observer's eye examines the prism in direction  $CO$ .

A simple photometer was proposed by the German physicist and chemist R. Bunsen (1811-1899). The optical field in this photometer is a screen of white paper with a greased transparent central region at the middle. The grease spot must have sharp edges. Two light sources are placed on both sides of the screen. By weakening the luminous flux from one of the sources, the grease spot and the remaining part of the screen can be made equally bright. Many other more perfect photometers are designed on the basis of this principle of "transparent zone".

A simple way to obtain equal illuminances of two elements of a photometer is to change the distances from the sources being compared to the photometer in the case when the inverse-square law is applicable (see Sec. 8.4). The illuminance of the surface element is proportional to the luminous intensity of the source and inversely proportional to its squared distance from the surface element. If the illuminances of the two elements are equal, we have

$$\frac{I_1}{R_1^2} = \frac{I_2}{R_2^2},$$

where  $I_1$  and  $I_2$  are the luminous intensities and  $R_1$  and  $R_2$  are the distances from the sources to the photometer. Having measured  $R_1$  and  $R_2$ , we can determine the ratio of the luminous intensities of the sources. This method has a drawback since the distances  $R_1$  and  $R_2$  can be practically varied only within narrow range.

Another way of weakening the luminous flux from one of the sources consists in that an absorbing body is placed on the path of light. This body can be made in the form of two wedges sliding relative to each other and made of a light-absorbing material (Fig. 166). By moving the wedges, we can vary the thickness of the absorbing layer, and hence the degree of absorption of the luminous flux. The attenuator is calibrated beforehand, for which purpose the change in absorption caused by a displacement of a wedge through a certain distance is determined.

There exist photometers intended for a direct measurement of illuminance. Such photometers are called *luxometers*.

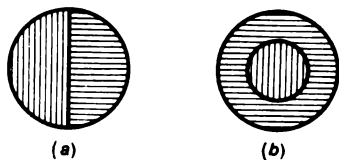


Fig. 164.

The types of fields compared in a photometer.

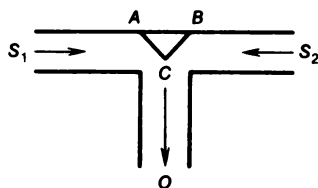


Fig. 165.

Schematic diagram of a simple photometer.

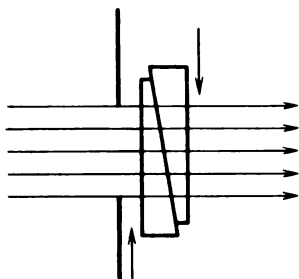


Fig. 166.

An appliance for weakening luminous flux, which ensures the passage of rays without a deflection.

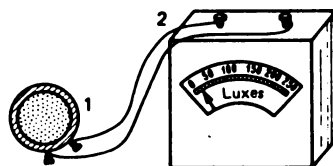


Fig. 167.

Luxometer: 1 — photocell, 2 — galvanometer with a scale graduated in luxes.

In luxometers, light is perceived by a *photocell*. The electric current emerging in a photocell under the action of light is the larger, the higher the illuminance of the photocell (naturally, provided that the surface of the photocell is illuminated uniformly). Therefore, the measurement of illuminance with the help of an objective photometer is reduced to the measurement of the current passing through a galvanometer connected with the photocell (see Sec. 20.10 for details).

Figure 167 shows schematically a *luxometer*. It consists of a photoelectric cell 1 connected to galvanometer 2 with the help of a wire. The galvanometer scale is graduated directly in luxes. In order to measure the illuminance with the help of this instrument, it is sufficient to put the photocell on the surface whose illuminance is to be determined and take a reading from the scale. Photoelectric luxometers are convenient in operation and make it possible to take measurements rapidly and easily.

Sometimes, a photocell and a galvanometer are enclosed in a single casing. Such luxometers are used by amateur photographers for determining the illuminance of an object to be photographed, and hence for choosing appropriately the exposure. For this reason, they are called exposure meters

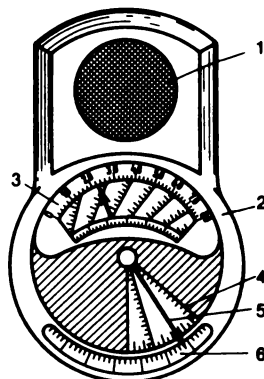


Fig. 168.

Exposure meter: 1 — photocell, 2 — galvanometer, 3 — scale of exposure time, 4 — scale of sensitivity of photographic materials, 5 — indicator, 6 — scale of diaphragm diameter.

(Fig. 168). The scale of the galvanometer of an exposure meter graduated for the duration of exposure is arranged on the semicircle of a rotating ring 3. The divisions on sector 4 rotating together with ring 3 correspond to the sensitivity of photographic materials to be used. Indicator 5 is arranged against the division on the stationary scale 6, corresponding to the diaphragm used during photographing. Then ring 3 is rotated until the required division on sector 6 coincides with indicator 5. The pointer of the galvanometer indicated the exposure time required for photographing with a given photographic material.

? II.1. What must be the power ratio of the blue radiation ( $\lambda = 460 \text{ nm}$ ) and the yellow-green radiation (corresponding to the maximum sensitivity of the human eye) for the visual sensation to be the same?

II.2. The inverse-square law cannot be used for calculating the illuminance from a large source. However, we can mentally divide the entire surface of a large source into small regions such that the inverse-square law is applicable to each of them. Then why is this technique of calculating illuminance inapplicable to the light source as a whole?

II.3. It was mentioned above that a parallel beam of light cannot be obtained experimentally. What is then meant by saying that the basic property of a lens is the formation of a parallel light beam when a source is at the lens focus?

II.4. What luminous flux is incident on a surface whose area is  $100 \text{ cm}^2$  at a bright sunny noon when the illuminance reaches  $100\,000 \text{ lx}$ ?

II.5. A luminous flux of  $10\,000 \text{ lm}$  is incident on a surface whose area is  $4 \text{ m}^2$ . Find the illuminance of this surface.

II.6. The luminous intensity of a light source is  $100 \text{ cd}$ . Determine the total luminous flux emitted by the source and the illuminance of a surface perpendicular to the direction of the rays and located at  $3 \text{ m}$  from the source.

II.7. In a novel "The Engineer Garin's Hyperboloid" by A. Tolstoy, a device of a huge destructive power, based on the concentration of radiant energy in a very narrow (parallel) light beam is described (the schematic diagram of the device is shown in Fig. 169). Analyze the operation of the device and explain why it cannot produce the effect ascribed to it by the author.

II.8. What is the brightness of the surface having a reflection coefficient of  $0.9$  if the illuminance is  $100\,000 \text{ lx}$ ?

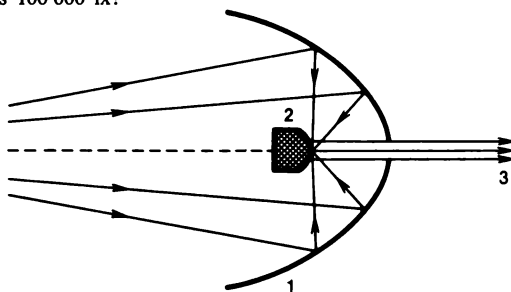


Fig. 169.

Schematic diagram of a "hyperboloid": 1 — converging hyperbolic mirror, 2 — reflecting hyperboloid, 3 — parallel bundle of rays.

- II.9.** Determine the brightness of a source emitting a luminous flux of 15 lm from  $1 \text{ mm}^2$  of its surface in a solid angle of 0.03 sr.
- II.10.** Determine the illuminance of the centre and the edge of a round table having a diameter of 3 m and illuminated by a lamp hanging at 2 m above its centre and having a luminous intensity of 200 cd.
- II.11.** A luminous flux of 2000 lm is incident on an opal glass. The reflected luminous flux is 500 lm. Determine the reflection and transmission coefficients and the absorbed and transmitted luminous fluxes if the absorption coefficient of the glass is 0.4.
- II.12.** A luminous flux of 1000 lm is incident on a chrome-coated reflecting surface. Find the reflected and absorbed luminous fluxes if the reflection coefficient for chrome is 0.65.
- II.13.** A luminous flux of 1000 lm is incident on a sheet of white paper whose area is  $500 \text{ cm}^2$ . The reflection coefficient of the paper is 0.68. Determine the illuminance and the brightness of the sheet of paper.
- II.14.** Determine the brightness of a snow field under solar rays producing an illuminance of 50 000 lx on it. The reflection coefficient of snow is 0.80.
- II.15.** The brightness of the Sun is  $1.5 \times 10^9 \text{ cd/m}^2$  and its diameter is  $1.4 \times 10^6 \text{ km}$ . Determine the luminous intensity of the Sun observed on the Earth and the illuminance of the terrestrial screen perpendicular to the solar rays. (The distance between the Sun and the Earth is  $1.5 \times 10^8 \text{ km}$ ).
- II.16.** Determine the axial brightness of the crater of an electric arc if the luminous intensity along its axis is 40 000 cd and the diameter is 15 mm.
- II.17.** The luminous intensity of a standard lamp is 25 cd. The distance from the standard lamp to the screen of photometer is 15 cm at the same brightness of compared fields. The distance from a tested lamp to the screen is 45 cm. Determine the luminous intensity of the tested lamp.



## Chapter 9

# Basic Laws of Geometrical Optics

### 9.1. Rectilinearity of Wave Propagation

To determine the mode of wave propagation, we must, strictly speaking, find out how a perturbation is transmitted from one point of the medium to another, how perturbations interact and what the final result of this perturbation is. Experiments show, however, that in most cases, *when the size of the portion of a wave under consideration is large in comparison with the wavelength*, a number of simple laws simplify the solution of the problem about the wave propagation.

Let us repeat the experiment with waves on water surface, caused by vibrations of the edge  $LL$  of a ruler disturbing the water surface. In order to determine the direction of wave propagation, we place on their way a screen  $MM$  with a hole whose size is considerably larger than the wavelength (as shown in Fig. 87a, Sec. 4.9). In accordance with what was said in Sec. 4.9, waves propagate behind the screen within a straight channel drawn through the edges of the hole (Fig. 170). The direction of this channel is just the direction of wave propagation. It remains unchanged if we arrange the screen *at an angle* ( $M'M'$ ). The direction along which the waves propagate always turns out to be *perpendicular* to the line all of whose points are reached by a wave perturbation at the same instant. This line is known as the *wave front*.<sup>1</sup> The straight line perpendicular to a wave front (the arrow in Fig. 170) indicates the direction of wave propagation. We shall refer to this line as a *ray*. Thus, *a ray is a geometrical line drawn at right angles to a wave front and indicating the direction of propagation of a wave perturbation*.

The perpendicular to a wave front, i.e. the ray, can be drawn at any point of the wave front.

In the case considered above, the wave front has the form of a straight line. Therefore, the rays are parallel to one another at all points of the wave front. If we repeat the experiment with the vibrating end of a wire as a source, the wave front will have the shape of a circle. Having placed on

---

<sup>1</sup> For waves propagating over water surface, the wave front is a line. For space waves (acoustic or optical), the wave front is the surface all of whose points are reached by a wave perturbation at the same instant (wave surface).

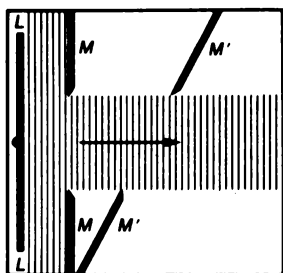


Fig. 170.

Waves behind a large aperture propagate within a straight channel drawn through the edges of the aperture.

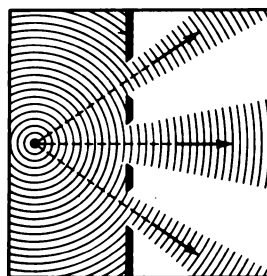


Fig. 171.

Propagation of waves through large apertures when a wave front has the circular form (the size of the apertures is reduced).

the way of such a wave screens with holes whose sizes are large in comparison with the wavelength, we obtain a pattern shown in Fig. 171. Thus, in this case too the direction of wave propagation coincides with the straight lines perpendicular to the wave front, i.e. with the direction of the rays. In the latter case, the rays are depicted as the radii emanating from the source of waves.

## 9.2. Rectilinear Propagation of Light.

### Light Rays

Observations show that in homogeneous media, light also propagates along straight lines. Some experiments illustrating this phenomenon are well known. When an object is illuminated by a *point source*, a sharp shadow (*umbra*) is formed (Fig. 172), whose shape is similar to the shape of a cross section of the object, which is parallel to the screen. The size of the umbra is determined by the mutual arrangement of the source, the object and the screen in conformity with the laws of projecting by straight lines. Partial

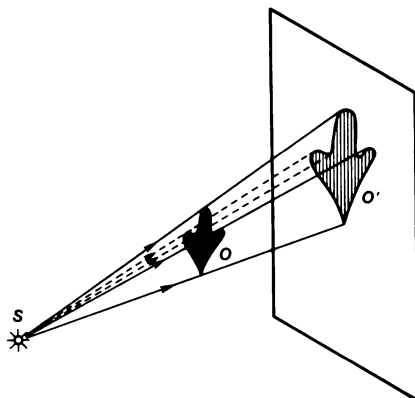
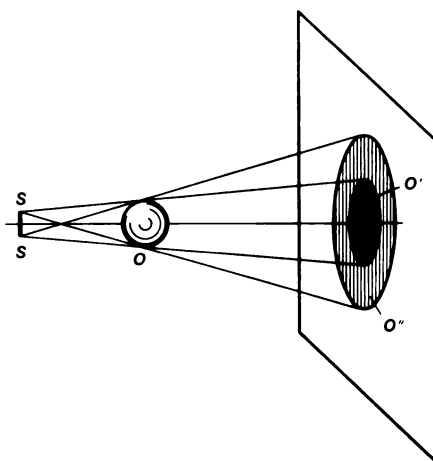


Fig. 172.

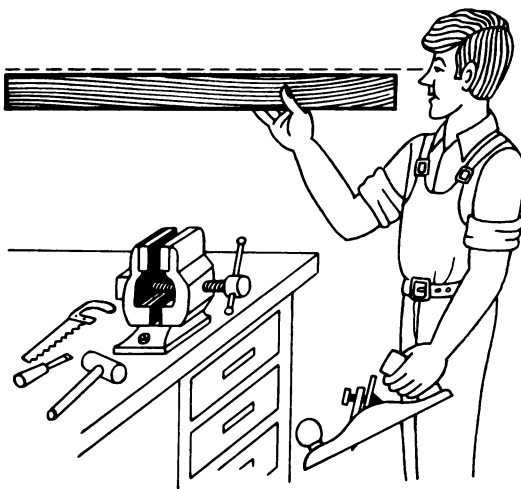
When a point source  $S$  illuminates a plane object  $O$  parallel to a screen, a sharp shadow (umbra)  $O'$  formed on the screen is similar to the object.



**Fig. 173.**

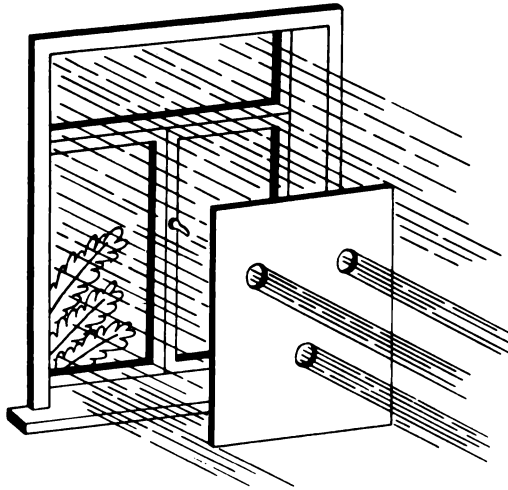
When ball  $O$  is illuminated by an extended source  $SS$ , the umbra  $O'$  on the screen is surrounded by penumbra  $O''$ .

shadows (penumbras) observed sometimes are due to the finite size of the light source rather than due to a deviation of the direction of propagation from a straight line (Fig. 173). Joiners employ a well-known method of verifying the straightness of a planed board “by ray” (Fig. 174). Phenomena associated with the rectilinear propagation of light are similar to those described in the previous section. If we make the paths of solar rays “visible” with the help of cigarette smoke, the experiment with partitions will be repeated for light. If we interpose a light ray with a cardboard screen having one or several small holes (which are naturally much larger than



**Fig. 174.**

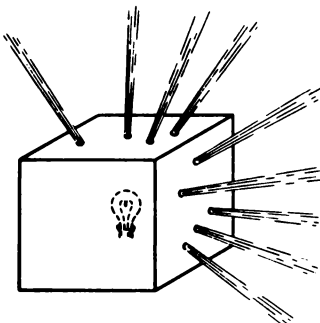
Verification of the straightness of the edge by the “line of sight”.

**Fig. 175.**

Beams isolated from a plane light wave (the light source is the Sun) by a screen with small holes. The diameters of the holes are large in comparison with the light wavelength.

the optical wavelength), we will see the path of light in the room in the form of narrow beams emanating from the holes (Fig. 175). With any position of the screen, these beams have the same direction pointing to the Sun.

If in a dark room we place a bright bulb with a small filament into a dark box with several holes, the path of emanating rays will be seen in the dusty air of the room as a narrow beam diverging in all directions (Fig. 176). Having marked the position of the bulb on the box wall, we will easily see that these beams intersect at the point where the bulb filament is located. Proceeding in the same way as in the case of the waves on water surface, we draw the lines indicating the direction of light propagation. If the isolated beams are narrow, the construction of such lines presents no difficulty. These geometrical lines are *light rays*. In the cases considered above, they

**Fig. 176.**

Beams of rays isolated from a spherical light wave.

are nearly parallel lines directed to the Sun, or the radii perpendicular to the surface of the sphere with the centre at the point where the light source (bulb filament) is located. A light wave propagates along these straight rays.

Sometimes, in school textbooks, light rays are identified with *narrow beams of light* used for determining the direction of the rays. This is wrong since a ray is a *geometrical line* indicating the direction of light propagation rather than of a light beam. Naturally, the narrower a light beam, the higher the accuracy with which the direction of light propagation, i.e. of a light ray is determined. However, an infinitely narrow light beam cannot be obtained.

Decreasing the size of the hole bounding a beam, we can reduce the beam width only to a certain limit. A further decrease in the hole size does not lead to the reduction in the cross-sectional area of the beam. On the contrary, experiments show that this leads to the divergence of the beam. In Sec. 4.9, this phenomenon was described for the waves on water surface (see Fig. 87*b* and *c*).

For light waves, this phenomenon can be observed while forming an image with the help of a small aperture (the so-called “aperture camera”<sup>2</sup>). These observations also show that the law of rectilinear propagation of light holds only under certain conditions. An experiment confirming this statement is represented in Fig. 177. An inverted image of a bright object (say, the filament of an incandescent lamp) located in front of a camera-obscura is formed on the frosted glass (or a photographic plate) covering the rear wall of the camera. The image reproduces the shape of the object and does not depend on the shape of the aperture if the latter is sufficiently small.

This result is quite understandable. Indeed, a narrow light beam from each point of the source passes through an aperture and forms on the screen a small spot reproducing the aperture shape. The light of the source as a whole produces on the screen a pattern formed by such bright spots. If the size of the aperture is such that individual spots are larger than the

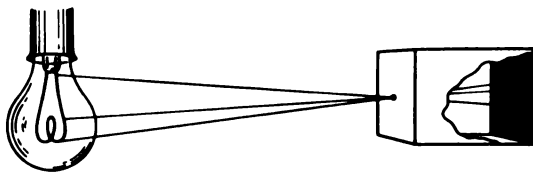


Fig. 177.

Image of an object formed in a pinhole camera. The size of the aperture is not indicated. Actually, a cone of rays corresponds to each ray, and for this reason the image of the bulb filament is slightly blurred.

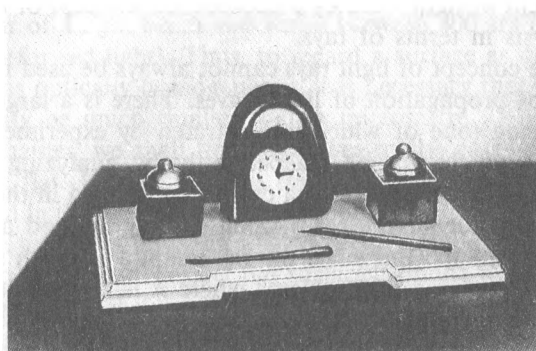
---

<sup>2</sup> This camera is also known as camera-obscura.

details of the pattern, the latter will be blurred and will represent the object inadequately. But if the size of the aperture is small enough, the spots will be smaller in size than the details, and the image will be quite satisfactory.

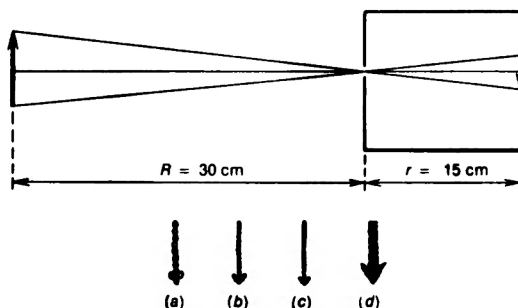
Figure 178 is a photograph obtained with the help of such a camera.

Figure 179 explains the operation of a camera-obscura and the form of the images obtained with different sizes of aperture. The improvement of the image due to the reduction of the aperture size is observed only to a certain limit. With a further decrease in the aperture size, the image is blurred (Fig. 179*d*). When the size becomes very small, the image does not resemble the source at all. This experiment indicates that bright spots representing individual points of the source are blurred (when the aperture



**Fig. 178.**

The photograph obtained with the help of a pinhole camera.



**Fig. 179.**

Schematic diagram of a pinhole camera (top) and images of a light source for different sizes of the apertures (bottom): (a) the aperture diameter is about 3 mm, (b) about 1 mm, (c) about 0.5 mm, (d) about 0.03 mm. The source is a brightly illuminated slit in a screen, having the form of an arrow with a width of about 1 mm.

size is very small) so that they exceed the details of the pattern, the blurring becoming stronger with a further reduction of the aperture size. But since these spots are the traces of light beams cut by the aperture, the experiment demonstrates the divergence of a light beam upon a considerable reduction of the aperture size. Therefore, it is physically impossible to isolate an infinitely narrow beam. We have to confine ourselves to the isolation of finite-width beams and then replace them by the lines representing the axes of these beams. Thus, *a light ray is a geometrical concept*.

The merit of this concept consists in the possibility of establishing the direction of propagation of luminous energy. The laws governing the change in the direction of rays allow us to solve very important optical problems concerning the change in the direction of propagation of luminous energy. For solving most of such problems, it is sufficient to replace a physical concept of light wave by a geometrical concept of light ray and carry out analysis in terms of rays.

However, the concept of light rays cannot always be used for determining the nature of propagation of light waves. There is a large number of optical phenomena (one of which is illustrated by experiments with the camera-obscura) which can be grasped only by analyzing light waves proper. Naturally, optical phenomena can be considered in the framework of the wave theory for simple cases when the ray method also provides quite satisfactory results. But since the ray method is much simpler, it is normally used for solving all problems where it is applicable. Therefore, we must outline the range of problems and the degree of accuracy with which geometrical rays can be used to distinguish these problems from those where the ray method leads to considerable errors and hence is inadmissible.

Thus, the method of *ray optics*, or *geometrical optics*, is an approximate technique for solving problems, which is adequate for analyzing a certain scope of questions. For this reason, it is important for those studying optics to learn how to apply the ray method correctly and to establish the limits of its applicability.

### 9.3. Laws of Reflection and Refraction of Light

As was mentioned above (see Sec. 8.9), we can see nonluminous objects due to the fact that any body partially reflects and partially transmits or absorbs light incident on it. In that section, we were interested mainly in diffuse reflection and transmission. It is due to these phenomena that light incident on a body is scattered in different directions, and we are able to see the body from any side.

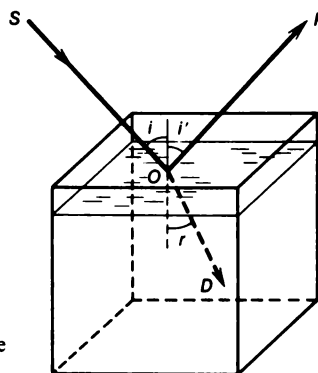
In particular, it is due to scattered (although weak) light that we see from any point very good mirrors which should reflect light only in one

direction and hence be seen only in one direction. Light is scattered in this case due to small defects on the surface, scratches, dust particles, and so on. In this chapter, we shall consider the *laws of direct (mirror) reflection and direct transmission (refraction) of light*.

In order to obtain mirror reflection or direct refraction, the surface of a body must be sufficiently smooth (not dull), and its internal structure must be sufficiently homogeneous (unturbid). This means that the unevenness of the surface or the inhomogeneity of the internal structure must be sufficiently small. As in any physical phenomenon, the expression “sufficiently small” or “sufficiently large” indicates a value small or large in comparison with some other physical quantity important for a phenomenon under investigation. In this case, this quantity is the *wavelength of light*. Henceforth, we shall indicate the methods of its measurement. Here, we shall confine ourselves to the remark that the wavelength of light depends on the colour of a light beam and varies between 400 nm (for violet light) and 760 nm (for red light). Thus, to regard a surface as optically smooth and a body as optically homogeneous, it is necessary that unevenness and inhomogeneity be much smaller than a micrometre.

In this chapter, we shall limit ourselves to the case when the surface of a body is flat. Passage of light through bent (spherical) surfaces will be considered in the next chapter. An example of a flat surface is the interface between air and a liquid in a wide vessel<sup>3</sup>. The appropriate polishing of solids also yields quite perfect flat surfaces, among which metal surfaces are distinguished by their ability to reflect much light. Glass is used for obtaining flat plates which can be coated with a metal layer to obtain conventional mirrors.

Let us consider the following simple experiment. We direct a narrow light beam on the water surface in a large vessel (Fig. 180). We shall see



**Fig. 180.**  
Refraction and reflection of light incident on the surface  
of water.

<sup>3</sup> In narrow vessels, the surface of liquid may be noticeably curved due to capillarity.



that light is partially reflected from the water surface, while some part of light will pass from air into water. To make the incident ray  $SO$ , reflected ray  $OR$  and transmitted ray  $OD$  visible, we recommend that air above the vessel be slightly dusted (say, by smoke) and a small amount of soap be dissolved in the water to make it slightly turbid. The experiment shows that the ray transmitted through the water is not a simple continuation of the ray incident on the interface but is refracted.

In the analysis of this phenomenon, we shall be interested, firstly, in *the direction of the reflected and refracted rays*, and secondly, in *the fraction of reflected luminous energy and the fraction of energy transmitted from the first to the second medium*.

Let us first consider reflected rays. We shall cover the interface (or a mirror) by an opaque cylindrical surface  $ACB$  made, for example, from a thick paper (Fig. 181a). Along arc  $ACB$ , we make small holes separated, say, by  $5^\circ$ . Then it will turn out that a ray passed through one such hole along the radius of arc  $ACB$  to its centre  $O$  will emerge after reflection through a hole symmetric about normal  $NO$  to the entrance. Irrespective of the accuracy with which this experiment is made, its result remains the same even if the most perfect instrument for measuring angles is used. This reliable result can be formulated as the following *law of reflection*: *the incident ray, the reflected ray and the normal to the reflecting surface lie in the same plane, the angle of reflection being equal to the angle of incidence*.

The angle formed by the refracted ray and the normal to the interface (*angle of refraction*) can be measured in the same way as that used above for the angle of reflection. For this purpose, the cylindrical surface  $ACB$  should be extended to the second medium (Fig. 181b). Accurate measurements of the angle of incidence  $i$  and the angle of refraction  $r$  lead to the following *law of refraction*: *the incident ray, the refracted ray and the normal to the interface lie in the same plane, the angle of incidence being*

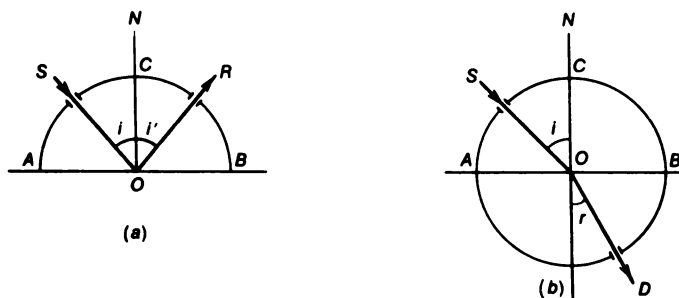


Fig. 181.

Measurement of the angle of reflection (a) and refraction (b).

connected with the angle of refraction through the following relation:

$$\frac{\sin i}{\sin r} = n, \quad (9.3.1)$$

where the refractive index  $n$  is a constant independent of the angle of incidence and determined by the optical properties of the media in contact.

The angles  $i$ ,  $i'$  and  $r$  of incidence, reflection and refraction are measured from the normal to the interface to the corresponding ray.

The first attempts to establish the law of refraction were made by the famous Alexandrian astronomer Claudius Ptolemy (died *circa* 168 A. D.) almost two thousand years back. However, the accuracy of measurement was not sufficiently high at that time, and Ptolemy drew the conclusion that the ratio of the angles of incidence and refraction was constant for given media. It should be noted that in order to obtain the correct dependence between the angle of incidence and the angle of refraction, these angles should be measured to within a few minutes. This is especially important for small angles of incidence and refraction. For rough measurements and small angles, the erroneous conclusion about a constant ratio of the angles instead of the constant ratio of the sines of the angles<sup>4</sup> can be drawn as was the case with Ptolemy. In the correct form, the law of refraction was established only one and a half thousand years later by the Dutch mathematician Willebrod Van Roijen Snell (1591-1626) and, apparently, by the French physicist and mathematician René Descartes (1596-1650) independently.

Let us now consider the *amount of reflected luminous energy*. The image of the face of a person in a good mirror is known to be brighter than on the water surface in a lake or well. This is due to the fact mentioned more than once that not the entire energy of light incident on the interface between two media is reflected from it: a part of light penetrates through the interface into the second medium and passes through it or is partially absorbed by it.

The fraction of the reflected luminous energy depends on the optical properties of contacting media and on the *angle of incidence*. If, for example, light is incident on a glass plate at right angles to its surface (the angle of incidence is zero), only about 5% of the luminous energy is reflected, while 95% passes through the interface. As the angle of incidence increases, the fraction of the reflected energy increases. Table 4 shows, by way of an example, the fraction of energy reflected at different angles of incidence of light on the air-glass interface ( $n = 1.555$ ). Table 5 contains similar data for the air-water interface ( $n = 1.333$ ).

Concluding the section, it should be stipulated that the laws of reflection and refraction are valid only if the interface is much larger than the

---

<sup>4</sup> Since for small angles  $\sin \alpha \approx \alpha$  (angle  $\alpha$  is expressed in radians),  $\sin i \approx i$ ,  $\sin r \approx r$ , and hence  $n = \sin i / \sin r \approx i / r$  (in the last relation, the angles can be expressed in degrees since the ratio of similar quantities does not depend on the choice of the unit of their measurement).

**Table 4.** Fraction of Reflected Energy at Various Angles of Incidence of Light on Glass Surface

| Angle of incidence                    | 0°   | 10°  | 20°  | 30°  | 40°  | 50°  | 60°  | 70° | 80° | 89° | 90° |
|---------------------------------------|------|------|------|------|------|------|------|-----|-----|-----|-----|
| Fraction of reflected energy (in %)   | 4.7  | 4.7  | 4.7  | 4.9  | 5.3  | 6.6  | 9.8  | 18  | 39  | 91  | 100 |
| Fraction of transmitted energy (in %) | 95.3 | 95.3 | 95.3 | 95.1 | 94.7 | 93.4 | 90.2 | 82  | 61  | 9   | 0   |

**Table 5.** Fraction of Reflected Energy at Various Angles of Incidence of Light on Water Surface

| Angle of incidence                    | 0°   | 10°  | 20°  | 30°  | 40°  | 50°  | 60°  | 70°  | 80°  | 89°  | 90° |
|---------------------------------------|------|------|------|------|------|------|------|------|------|------|-----|
| Fraction of reflected energy (in %)   | 2.0  | 2.0  | 2.1  | 2.2  | 2.5  | 3.4  | 6.0  | 13.5 | 34.5 | 90.0 | 100 |
| Fraction of transmitted energy (in %) | 98.0 | 98.0 | 97.9 | 97.8 | 97.5 | 96.6 | 94.0 | 86.5 | 65.5 | 10.0 | 0   |

wavelength of light. For example, a small mirror acts as a small aperture, the only difference being that the mirror changes the direction of rays incident on it. If the mirror size is smaller than 0.01 mm, the wave properties of light become noticeable just as for the passage of light through very small apertures. In this case, a narrow beam of light diverges upon reflection, the divergence being the stronger, the smaller the mirror size. The same applies to the refracted beam. The explanation of these phenomena will be given in the chapter devoted to diffraction of light.

**9.4. Reversibility of Light Rays**

While analyzing in the previous section the phenomena observed when light is incident on the interface between two media, we assumed that light propagates in a certain direction indicated by arrows in Figs. 180 and 181. Let us now see what happens when light propagates in the reverse direction. For the case of reflection, this means that the incident ray is directed not from the left and downwards as in Fig. 182*a*, but from the right and down-

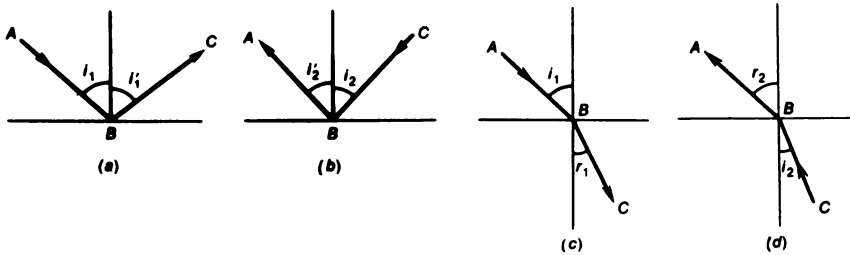


Fig. 182.

Reversibility of light rays in reflection (a, b) and in refraction (c, d). If  $i_2 = r_1$ , then  $r_2 = i_1$ .

wards as in Fig. 182b. For the case of refraction, we shall consider the propagation of light not from the first to the second medium as in Fig. 182c, but from the second to the first medium as in Fig. 182d.

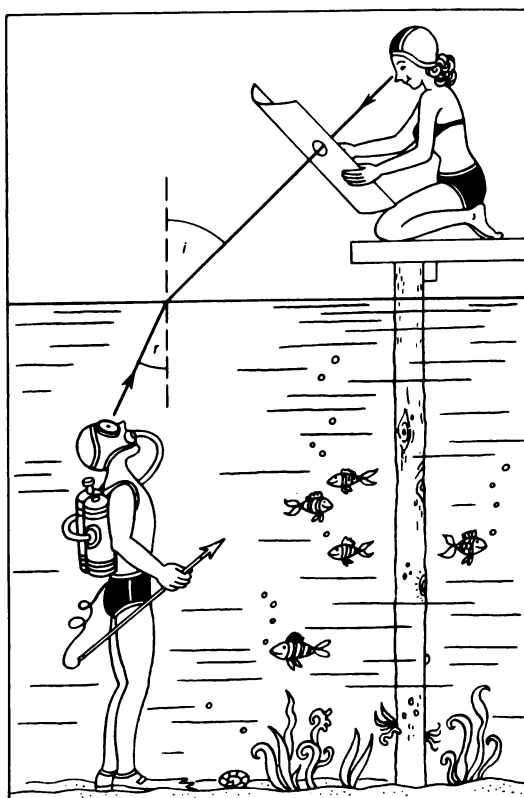
Accurate measurements show that both for reflection and refraction, the angles between the rays and the normal to the interface remain unchanged, and only the direction of arrows is reversed. Therefore, if a light ray is incident along  $CB$  (see Fig. 182b), the reflected ray will propagate along  $BA$ , i.e. the *incident and reflected rays have changed places* in comparison with the former case. The same is observed for refraction of a light ray. Let  $AB$  be an incident ray and  $BC$  a refracted ray (see Fig. 182c). If the light is incident along  $CB$  (see Fig. 182d), the refracted ray will propagate along  $BA$ , i.e. the *incident and refracted rays have changed places*.

Thus, both in reflection and refraction, light can follow the same path in the opposite directions (Fig. 183). This property of light is known as the *reversibility of light rays*.

The reversibility of light rays indicates that if the refractive index is  $n$  for the transition from the first to the second medium for the reverse transition it will be  $1/n$ . Indeed, let the light be incident at angle  $i$  and refracted at angle  $r$  so that  $n = \sin i / \sin r$ . If for the reverse direction of the rays, the light is incident at angle  $r$ , it must be refracted at angle  $i$  (reversibility). In such a case, the refractive index  $n' = \sin r / \sin i$ , and hence  $n' = 1/n$ . For example, when light passes from air to glass,  $n = 1.50$ , while for the reverse transition from glass to air,  $n' = 0.67 = 1/1.50$ .

The reversibility of light rays is preserved in multiple reflections and refractions which may take place in any order. This follows from the fact that in each reflection or refraction, the direction of a light ray can be reversed.

Thus, *if a light ray emerging from any system of reflecting and refracting media is reflected at the last stage exactly in the backward direction, it will pass through the entire system in the reverse direction and return to its source.*

**Fig. 183.**

To the reversibility of light rays in refraction.

The reversibility of light rays can be proved theoretically from the laws of reflection and refraction without resorting to new experiments. In the case of reflection, the proof is very simple (see Ex. II.22 at the end of the chapter). A more complicated proof for the case of refraction can be found in textbooks on optics.

### 9.5. Refractive Index

Let us consider in greater detail the concept of refractive index introduced in Sec. 9.3 while formulating the law of refraction.

Refractive index depends on the optical properties of the medium into which light penetrates. The refractive index obtained when light gets from vacuum into a medium is known as the *absolute refractive index for a given medium*.

Let the absolute refractive index of the first medium be  $n_1$  and of the

second medium  $n_2$ . Considering the refraction at the interface between the first and second media, we can make sure that the refractive index  $n$  for the transition from the first to the second medium (*relative refractive index*) is equal to the ratio of the absolute refractive indices of the second and first media:

$$n = \frac{n_2}{n_1} \quad (9.5.1)$$

(Fig. 184). On the contrary, when the light propagates from the second to the first medium, the relative refractive index is

$$n' = \frac{1}{n} = \frac{n_1}{n_2}. \quad (9.5.2)$$

These relations between the relative refractive index for two media and their absolute refractive indices could be derived theoretically without new experiments in the same way as it can be done for the reversibility principle (see Sec. 9.4).

A medium having a larger refractive index is referred to as an *optically denser medium*. The refractive index of various media is normally determined relative to air. The absolute refractive index of air  $n_{\text{air}}$  is 1.003. Thus, the absolute refractive index  $n_{\text{abs}}$  of a medium is connected with its refractive index relative to air ( $n_{\text{rel}}$ ) through the formula

$$n_{\text{abs}} = n_{\text{air}} \cdot n_{\text{rel}} = 1.003 n_{\text{rel}}. \quad (9.5.3)$$

Table 6 contains relative refractive indices determined for some cases of refraction of light at the interface between air and a corresponding medium.

Refractive index depends on the wavelength of light, i.e. on its *colour*. Different refractive indices correspond to different colours. This phenomenon, which is known as *dispersion*, plays an important role in optics. We

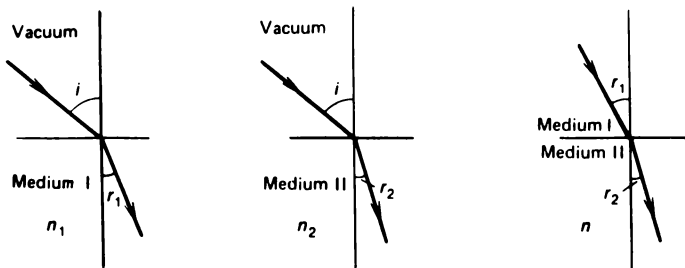


Fig. 184.

Relative refractive index for two media:

$$n_1 = \frac{\sin i}{\sin r_1}, \quad n_2 = \frac{\sin i}{\sin r_2}, \quad n = \frac{\sin r_1}{\sin r_2} = \frac{n_2}{n_1}.$$

**Table 6.** Refractive Indices of Some Substances Relative to Air

| Liquid            |       | Solid                |       |
|-------------------|-------|----------------------|-------|
| Substance         | $n$   | Substance            | $n$   |
| Carbon disulphide | 1.632 | Diamond              | 2.417 |
| Ethyl alcohol     | 1.362 | Glass (heavy flint)* | 1.80  |
| Glycerine         | 1.47  | Glass (light crown)* | 1.57  |
| Liquid helium     | 1.028 | Ice                  | 1.31  |
| Liquid hydrogen   | 1.12  | Ruby                 | 1.76  |
| Water             | 1.333 | Sugar                | 1.56  |

\* Crown and flint are the grades of optical glass

shall often deal with this phenomenon in the following chapters. The data contained in Table 6 refer to yellow light.

It is interesting to note that the law of reflection can formally be written in the same form as the law of refraction. It should be recalled that we agreed to measure angles from the normal to the corresponding ray. Consequently, the angle of incidence  $i$  and the angle of reflection  $i'$  should be taken with opposite signs. Then the law of reflection should be written in the form

$$i' = -i$$

or

$$\frac{\sin i}{\sin i'} = -1. \quad (9.5.4)$$

Comparing this formula with the law of refraction, we see that the law of reflection can be treated as a special case of the law of refraction for  $n = -1$ . This formal similarity of the laws of reflection and refraction is very useful for solving practical problems.

In the above analysis, refractive index was regarded as a constant characterizing a medium and independent of the intensity of light propagating through it. Such an interpretation of refractive index is quite natural, but it is unjustified for high intensities of radiation attained with the help of modern lasers. In this case, the properties of a medium through which a high-power optical radiation propagates depend on the radiation intensity. The medium is said to become nonlinear. The nonlinearity of the medium is manifested, in particular, in the fact that a high-intensity light wave changes the refractive index in it. The dependence of the refractive index on the radiation intensity  $J$  has the form

$$n = n_0 + aJ.$$

Here  $n_0$  is an ordinary refractive index and  $aJ$  is the nonlinear refractive index, where  $a$  is the proportionality factor. The additional term in the formula can be either positive or negative.

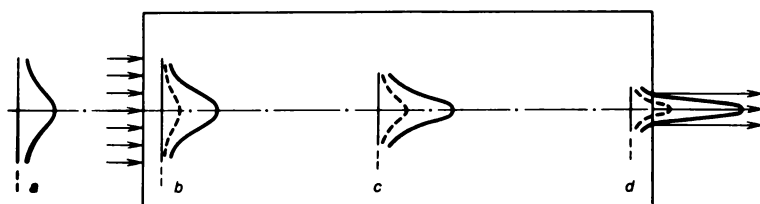


Fig. 185.

Distribution of radiation intensity and refractive index over the laser beam cross section at the entrance of the cuvette (*a*), near the input endface (*b*), in the middle of the cuvette (*c*), and near the output endface of the cuvette (*d*).

The relative variations of the refractive index are comparatively small. For  $J = 10^{12} \text{ W/m}^2$ , the nonlinear refractive index  $nJ = 10^{-5}$ . However, even such small variations of refractive index are noticeable: they are manifested in a peculiar effect of self-focussing of a light beam.

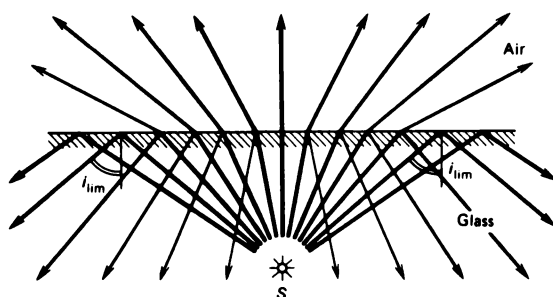
Let us consider a medium with a positive nonlinear refractive index. In this case, the regions of a higher intensity of light are at the same time the regions of an increased refractive index. In a real laser radiation, the intensity distribution over the cross section of the beam is usually nonuniform: the intensity is maximum on the axis and smoothly drops to the beam boundaries (solid curves in Fig. 185). Such a distribution also describes the variation of the refractive index over the cross section of the cuvette containing the nonlinear medium with a laser beam propagating along its axis. The refractive index is maximum on the cuvette axis and smoothly decreases towards its walls (dashed curves in Fig. 185).

While penetrating a medium with a refractive index  $n$ , the beam of rays emitted by a laser parallel to the axis is deflected towards higher values of  $n$ . Therefore, the higher intensity near the cuvette axis leads to the concentration of light rays in this region which is shown schematically in sections *b* and *c* in Fig. 185. This leads to a further increase in  $n$ . Ultimately, the effective cross section of the light beam passing through the nonlinear medium considerably decreases. The light passes as if along a narrow channel with a higher refractive index. Therefore, the laser beam contracts, and the nonlinear medium operates under the effect of high-intensity radiation as a converging lens. This phenomenon is known as *self-focussing*. It can be observed, for example, in liquid nitrobenzene.

## 9.6. Total Internal Reflection

It was mentioned in Sec. 9.3 that when light is incident on the interface between two media, the luminous energy splits into two parts: one part is reflected, while the other part penetrates through the interface into the second medium. Considering the example of light propagation from air to glass, i.e. towards the optically denser medium, we learned that the frac-



**Fig. 186.**

Total internal reflection: the thickness of rays corresponds to the fraction of the reflected luminous energy or that transmitted through the interface.

tion of reflected energy depends on the angle of incidence. In this case, the fraction of reflected energy sharply increases with the angle of incidence. However, even at very large angles of incidence close to  $90^\circ$ , when the light ray almost slips along the interface, a part of the luminous energy still goes to the second medium (see Sec. 9.3, Tables 4 and 5).

A new interesting phenomenon is observed if light propagating in a medium is incident on the interface between this medium and an optically less dense medium, i.e. one having a smaller absolute refractive index. In this case, the fraction of the reflected energy increases with the angle of incidence, but this increase obeys another law: *starting from a certain value of the angle of incidence, the entire luminous energy is reflected from the interface*. This phenomenon is known as the *total internal reflection*.

Let us consider again, as in Sec. 9.3, the incidence of light on the glass-air interface. Let a light ray be incident from glass on the interface at different angles (Fig. 186). If we measure the fraction of the reflected luminous energy and the fraction of the luminous energy transmitted through the interface, we obtain the values compiled in Table 7 (as in Table 4, the refractive index of glass is  $n = 1.555$ ). The value  $i_{cr}$  of the angle of incidence starting from which the entire luminous energy is reflected from the interface is called the *critical angle of total internal reflection*. For the grade of glass for which Table 7 is compiled ( $n = 1.555$ ), the critical angle is approximately equal to  $40^\circ$ .

**Table 7.** Fraction of Reflected Energy  
for Various Angles of Incidence of Light  
Propagating from Glass to Air

| Angle of incidence $i$              | $0^\circ$ | $10^\circ$     | $20^\circ$ | $30^\circ$ | $35^\circ$ | $38^\circ$     | $39^\circ$ | $39^\circ 30'$ | $40^\circ$ | $50^\circ$ | $60^\circ$ | $70^\circ$ | $80^\circ$ |
|-------------------------------------|-----------|----------------|------------|------------|------------|----------------|------------|----------------|------------|------------|------------|------------|------------|
| Angle of refraction $r$             | $0^\circ$ | $15^\circ 40'$ | $32^\circ$ | $51^\circ$ | $63^\circ$ | $73^\circ 20'$ | $79^\circ$ | $82^\circ$     | $90^\circ$ | —          | —          | —          | —          |
| Fraction of reflected energy (in %) | 4.7       | 4.7            | 5.0        | 6.8        | 12         | 23             | 36         | 47             | 100        | 100        | 100        | 100        | 100        |

Pay attention to the fact that when light is incident on the interface at the critical angle, the angle of refraction is  $90^\circ$ . In other words, in the formula

$$\frac{\sin i}{\sin r} = \frac{1}{n}$$

expressing the law of refraction for the case under consideration, we must put  $r = 90^\circ$  or  $\sin r = 1$  for  $i = i_{cr}$ . Hence we obtain

$$\sin i_{cr} = \frac{1}{n}. \quad (9.6.1)$$

For angles of incidence exceeding  $i_{cr}$ , the ray is not refracted at all. Formally, this follows from the fact that for angles of incidence greater than  $i_{cr}$ , the values obtained for  $\sin r$  from the law of refraction are greater than unity, which is obviously impossible.

Table 8 contains the critical angles of total internal reflection for some substances whose refractive indices are given in Table 6. The validity of relation ((9.6.1) follows from this table.

**Table 8.** Critical Angle of Total Internal Reflection at the Boundary with Air (in degrees)

| Substance         | $i_{cr}$ | Substance           | $i_{cr}$ |
|-------------------|----------|---------------------|----------|
| Carbon disulphide | 38       | Diamond             | 24       |
| Glycerine         | 43       | Glass (heavy flint) | 34       |
| Water             | 49       | Glass (light crown) | 40       |

Total internal reflection can be observed at the boundaries of air bubbles in water. They shine since the solar rays incident on them are completely reflected, without penetrating the bubbles. This is especially noticeable for the air bubbles which are always present on pedicles and leaves of seaweed and which, when observed in solar rays, appear as if made of silver, i.e. are a good reflector of light.

Total internal reflection is used in the construction of glass *reflecting and erecting prisms*. The operation of these prisms is explained by Fig. 187. The critical angle for a prism is  $35\text{--}40^\circ$  depending on the refractive index of a given grade of glass. Therefore, the angles of incidence and emergence of light rays can easily be selected for such prisms. Reflecting prisms successfully play the role of mirrors and are preferable since their reflecting properties remain unchanged, while metal mirrors grow dull with time because of the oxidation of metals. It should be noted that an erecting prism is simpler in construction than any erecting mirror system. Erecting prisms are used, in particular, in periscopes.

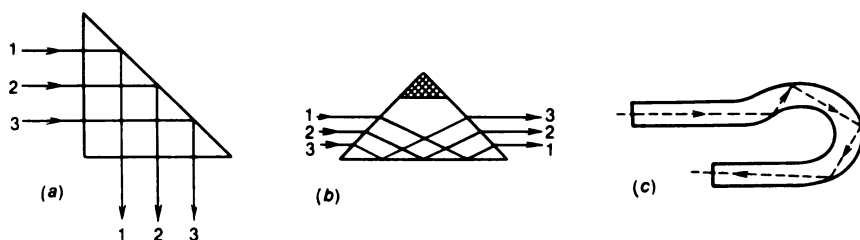


Fig. 187.

Path of the rays in a glass reflecting prism (a), erecting prism (b) and in a bent plastic tube—light guide (c).

### 9.7. Refraction in a Plane-parallel Plate

Let ray  $AB$  (Fig. 188) be incident on a plane-parallel glass plate. Having been refracted by glass, it will propagate in direction  $BC$ . At point  $C$ , it will undergo refraction again and will emanate from the plate in direction  $CD$ . Let us prove that ray  $CD$  emerging from the plate is parallel to the incident ray  $AB$ .

For the refraction at point  $B$ , we have

$$\frac{\sin i}{\sin r} = n,$$

where  $n$  is the refractive index of the plate. For the refraction at point  $C$ , the law of refraction gives

$$\frac{\sin r}{\sin i_1} = \frac{1}{n}$$

since in this case the ray emerges from the plate to air. Multiplying these expressions, we obtain

$$\sin i = \sin i_1.$$

or, since  $i \leq 90^\circ$  and  $i_1 \leq 90^\circ$ ,

$$i = i_1,$$

whence it follows that rays  $AB$  and  $CD$  are parallel.

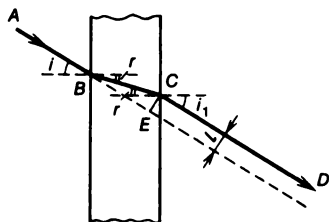


Fig. 188.

Refraction in a plane-parallel plate.

Ray  $CD$  is shifted relative to the incident ray  $AB$ . The shift  $l = EC$  depends on the thickness of the plate and on the angles of incidence and refraction. The shift is obviously the smaller, the thinner the plate.

### 9.8. Refraction in a Prism

Let ray  $AB$  be incident on a face of a prism. Having been refracted at point  $B$ , the ray propagates along  $BC$ , and after the second refraction at point  $C$ , emerges from the prism to air (Fig. 189). Let us find the angle  $\alpha$  through which the ray is deflected from its initial direction as a result of refraction in the prism. This angle will be referred to as the *angle of deflection*. The angle between the refracting faces of the prism, known as the *apical*, or *refracting, angle of the prism*, will be denoted by  $\theta$ . From quadrangle  $BOCN$  in which angles  $B$  and  $C$  are right angles, we find that angle  $BNC$  is  $180^\circ - \theta$ . Then from quadrangle  $BMCN$  we obtain

$$(180^\circ - \alpha) + (180^\circ - \theta) + i + i_1 = 360^\circ,$$

whence

$$\alpha = i + i_1 - \theta. \quad (9.8.1)$$

Angle  $\theta$ , being the external angle of triangle  $BCN$ , is given by

$$\theta = r + r_1, \quad (9.8.2)$$

where  $r$  is the angle of refraction at point  $B$  and  $r_1$  is the angle of incidence of the ray emerging from the prism at point  $C$ . Further, using the law of refraction, we have

$$\sin i = n \sin r, \quad (9.8.3)$$

$$\sin i_1 = n \sin r_1. \quad (9.8.4)$$

Knowing the apical angle  $\theta$  of the prism and the refractive index  $n$ , we can calculate the angle of deflection  $\alpha$  for any angle of incidence  $i$  with the help of obtained equations.

The expression for the angle of deflection assumes an especially simple form when the apical angle  $\theta$  of the prism is small, i.e. the prism is thin, and the angle of incidence  $i$  is not large. Then angle  $i_1$  is also small. Sub-

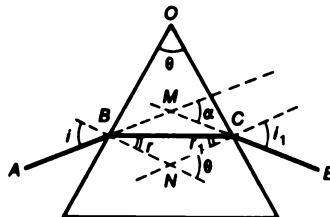


Fig. 189.  
Refraction in a prism.

stituting the approximate values of the angles (in radians) for the sines of these angles in (9.8.3) and (9.8.4), we obtain

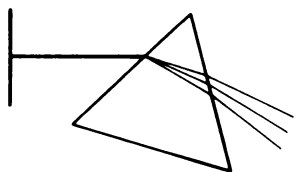
$$i = nr, \quad i_1 = nr_1.$$

Substituting these expressions into (9.8.1) and using (9.8.2), we get

$$\alpha = n(r + r_1) - \theta = (n - 1)\theta. \quad (9.8.5)$$

We shall use in the further discussion this formula which is valid for a thin prism at small angles of incidence.

Pay attention to the fact that the angle of deflection of a ray in a prism depends on the refractive index of the substance of which the prism is made. It was mentioned above that refractive indices are different for different colours (dispersion). For transparent bodies, the refractive index is maximum for violet rays, followed by indigo, blue, green, yellow, orange, and finally red rays for which the refractive index is minimum. Accordingly, the angle of deflection  $\alpha$  is maximum for violet rays and minimum for red rays. Therefore, a beam of white light incident on a prism turns out to be decomposed into a series of coloured rays (Fig. 190 and Fig. I on the fly-leaf), i.e. forms the *spectrum* of rays.



**Fig. 190.**

Dispersion of white light during the refraction in a prism. Incident beam of white light is shown as a front perpendicular to the direction of propagation of the wave. For refracted beams, only the direction of propagation is indicated.

? **II.18.** An image of a source can be obtained on a screen if a piece of cardboard with a hole in it is arranged in front of the screen. Under which conditions will the image on the screen be sharp? Explain why the image is inverted.

**II.19.** Prove that a beam of parallel rays remains parallel after the reflection at a plane mirror.

**II.20.** What is the angle of incidence of a ray if the incident and reflected rays make an angle of  $90^\circ$ ?

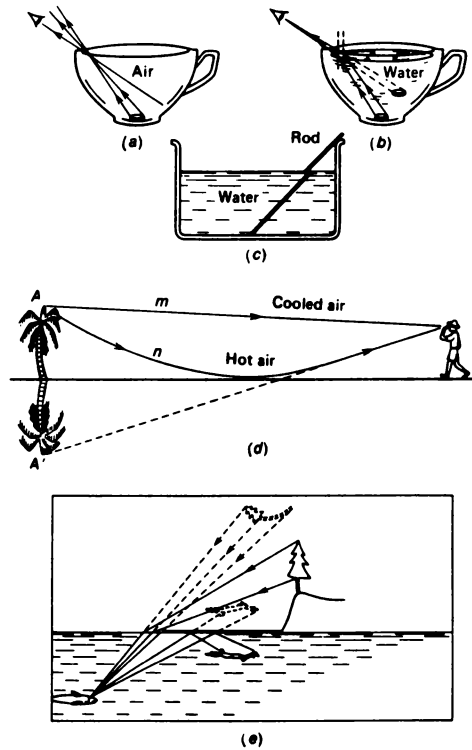
**II.21.** What is the angle of incidence if the reflected and refracted rays make an angle of  $90^\circ$ ? The refractive index of the second medium relative to the first is  $n$ .

**II.22.** Prove the reversibility of light rays in reflection.

**II.23.** Can a system of mirrors and prisms be designed so that the first observer will see the second observer, but the second observer will not see the first?

**II.24.** The refractive index of glass relative to water is 1.105. Determine the refractive index of glass relative to glycerine.

**II.25.** Find the critical angle of total internal reflection for the diamond-water interface.

**Fig. 191.**

To Exercise II.27. When the cup is empty, the eye does not see the coin (*a*). When the cup is filled with water, the coin becomes visible (*b*). A stick immersed in water by half looks as broken (*c*). A mirage in a desert (*d*). How a fish sees a tree and a diver (*e*).

**II.26.** Determine the shift of a ray passing through a plane-parallel plate made of a glass with a refractive index of 1.55 if the angle of incidence is  $45^\circ$  and the thickness of the plate is 1 cm.

**II.27.** Using the laws of reflection and refraction, explain the phenomena shown in Fig. 191.

## Chapter 10

# Application of Reflection and Refraction of Light for Image Formation

### 10.1. Light Source and Its Image

In Chap. 9, a general review of the laws of light individual laws and their applications of practical importance. In this chapter, we shall deal with refraction of light rays in a lens and reflection of rays from mirrors of various types.

It is known from everyday experience that looking at an object that serves as a source of light, we can get an idea about the location of this object. For solving similar problems, it is sufficient to trace any two rays emanating from a given element of a luminous object: the point of their intersection determines the location of a point source or, if the source is extended, a small element of this source. There is no need to consider other rays since all of them emanate from the same point and give nothing new for determining the position of this point.

The ability to correctly determine the location of luminous objects is acquired by man gradually, as a result of experience. A baby, for instance, tries to “catch” a star or the Sun by stretching hands in their direction. Gaining experience, a grown-up person can correctly estimate the distance to the objects emitting light.

In all cases when a point  $S'$  is the point of intersection and subsequent divergence of light rays, an eye (or any other receiver capable to react to the action of light) perceives these rays as if a light source were actually located at point  $S'$ . Such points at which light rays emanating from a real light source are focussed in one way or another are termed the *images* of this source (Fig. 192). The position of the image can be determined by constructing the paths of any two rays passing through it.

The images of point sources considerably differ from the actual point sources considered in Chap. 8 in that the rays propagate from them within a bounded solid angle, while they diverge uniformly in all directions from a real point source (cf. points  $S$  and  $S'$  in Fig. 192). For this reason, unlike a point source, an image cannot be seen from any point. This difference, however, is not of primary importance for this chapter, but is significant when the illuminance or brightness of image is considered (see Chap. 11).

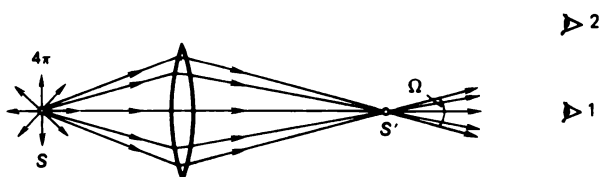


Fig. 192.

A real point source  $S$  can be seen from any position. Its image  $S'$  can be seen only within a bounded solid angle  $\Omega$  (position 1). In position 2, the image is invisible.

The formation of images of luminous points, as well as of extended objects, is the main problem in geometrical optics. Applying the laws of reflection and refraction, we shall primarily be interested in image formation.

### 10.2. Refraction in a Lens. Focal Points

In Chap. 9, we formulated the law of refraction indicating how the direction of a light ray changes when the light propagates from one medium to another. The simplest case of refraction of light at a plane interface between two media was considered.

For practical applications, of great importance is the refraction of light at a spherical interface. The main part of optical instruments, viz. a *lens*, is normally a glass body bounded on two sides by spherical surfaces. In particular, one surface of a lens can be a plane which can be regarded as a spherical surface of an infinitely large radius.

Lenses can be made not only of glass, but in general of any transparent material. In some instruments, lenses made of quartz, rock salt, etc. are used. It should be noted that the shape of the surfaces of lenses can be more intricate (for example, cylindrical or parabolic). Such lenses, however, are not widely used. Henceforth, we shall limit ourselves to the investigation of lenses with spherical surfaces.

Thus, let us consider a lens bounded by two spherical refracting surfaces  $PO_1Q$  and  $PO_2Q$  (Fig. 193). The centre of the first refracting surface lies

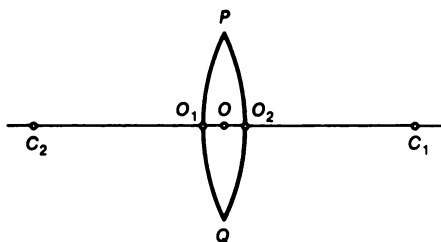


Fig. 193.

Thin lens:  $O$  — optical centre,  $C_1$  and  $C_2$  are the centres of spherical surfaces confining the lens.



at point  $C_1$ , while the centre of the second surface is at point  $C_2$ . For the sake of clarity, Fig. 193 shows a lens of a noticeable thickness  $O_1O_2$ . In actual practice, we shall usually assume that lenses under investigation are very thin, i.e.  $O_1O_2$  is very small in comparison with  $O_1C_1$  or  $O_2C_2$ . In such case, it can be assumed that points  $O_1$  and  $O_2$  practically merge into one point  $O$ . This point is known as the *optical centre* of the lens.

Every straight line passing through the optical centre of a lens is referred to as its *optical axis*. The axis passing through the centres of the two refracting surfaces is known as the *principal optical axis*, while all other axes are called *auxiliary optical axes*.

A ray propagating along an optical axis practically does not change its direction. Indeed, for rays passing along an optical axis, the regions of both surfaces of the lens can be regarded as parallel, and the thickness of the lens is assumed to be very small. It is well known that a light ray passing through a plane-parallel plate undergoes a parallel shift, but the shift of the ray in a very narrow plate can be neglected (see Ex. II.26).

If a ray is incident on a lens not along an optical axis but in any other direction, it will be deflected from the original direction as a result of the refraction first at the first and then at the second surface bounding the lens.

Let us cover a lens by a sheet of black paper 1 with an aperture exposing a small region in the vicinity of the principal optical axis of the lens (Fig. 194). We assume that the size of the aperture is small in comparison with  $O_1C_1$  and  $O_2C_2$ . We direct a parallel beam of light to lens 2 on the principal optical axis. The rays passing through the exposed part of the lens will be refracted and pass through a certain point  $F'$  lying on the principal optical axis to the right of the lens, at a distance  $f'$  from the optical centre  $O$ . If we place a white screen 3 at point  $F'$ , the region of intersection of the rays will have the form of a bright spot. The point  $F'$  on the principal optical axis, where the rays parallel to the principal optical axis intersect after the refraction in the lens, is called the *principal focus*, while the distance  $f' = OF'$  is known as the *focal length* of the lens.

Using the laws of refraction, it can easily be shown that all rays parallel to the principal optical axis and passing through a small central region of the lens indeed intersect after the refraction at a single point called the principal focus.

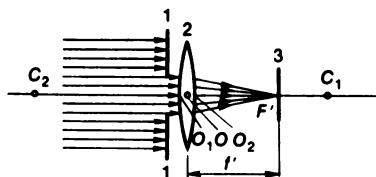


Fig. 194.  
Principal focus of a lens.

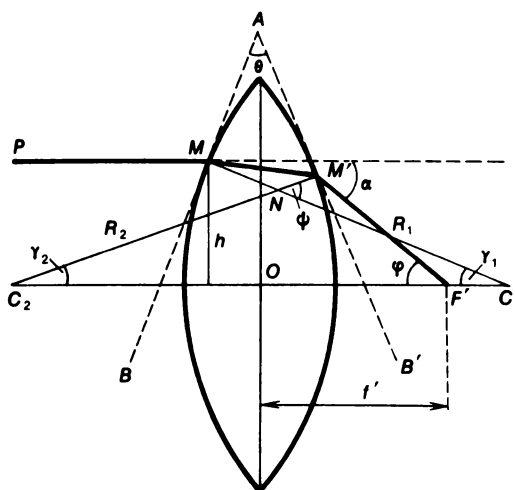


Fig. 195.

Refraction of ray  $PM$  parallel to the principal optical axis in a lens. (The lens thickness and height  $h$  are increased in comparison with distances  $R_1$ ,  $R_2$  and  $f'$ . Accordingly, angles  $\gamma_1$ ,  $\gamma_2$  and  $\theta$  in the figure are exaggerated.)

Let us consider ray  $PM$  incident on a lens parallel to its principal optical axis. Let this ray meet the first refracting surface of the lens at point  $M$  at a height  $h$  above the principal axis, the distance  $h$  being much smaller than  $C_2O$  and  $C_1O$  (Fig. 195). The refracted ray will propagate along  $MM'$  and after the refraction at the second surface bounding the lens will emerge from it along  $M'F'$ , making angle  $\varphi$  with the axis. The point of intersection of this ray and the axis will be denoted by  $F'$ , and the distance from this point to the optical centre of the lens, by  $f'$ .

Let us draw through points  $M$  and  $M'$  the planes tangential to the refracting surfaces of the lens. These planes (which are perpendicular to the plane of the figure) intersect at a certain angle  $\theta$  which is very small since the lens under consideration is thin. Instead of the refraction of ray  $PMM'F'$  in the lens, we can obviously consider the refraction of the same ray in a thin prism  $BAB'$  formed by tangential planes drawn through points  $M$  and  $M'$ .

It was shown in Sec. 9.8 that the angle through which a ray is deflected from the original direction in a thin prism is given by

$$\alpha = (n - 1)\theta, \quad (10.2.1)$$

where  $n$  is the refractive index of the substance of which the prism is made. Obviously, angle  $\alpha$  is equal to  $\varphi$  (see Fig. 195), i.e.

$$\varphi = \alpha = (n - 1)\theta. \quad (10.2.2)$$

Let  $C_1$  and  $C_2$  be the centres of the spherical refracting surfaces of the lens, and  $R_1$  and  $R_2$  are the corresponding radii of these surfaces. Radius  $C_1M$  is perpendicular to the tangential plane  $AB$ , and radius  $C_2M'$  is perpendicular to the tangential plane  $AB'$ . According to the well-known theorem from geometry, the angle between these perpendiculars, denoted by  $\psi$ , is equal to the angle  $\theta$  between the planes:

$$\psi = \theta. \quad (10.2.3)$$

On the other hand, angle  $\psi$  is equal to the sum of the angles  $\gamma_1$  and  $\gamma_2$  made by radii  $R_1$  and  $R_2$  and the axis as an external angle in triangle  $C_1NC_2$ :

$$\psi = \gamma_1 + \gamma_2. \quad (10.2.4)$$

Thus, using (10.2.2)-(10.2.4), we obtain

$$\varphi = (n - 1)(\gamma_1 + \gamma_2). \quad (10.2.5)$$

It was assumed that  $h$  is small in comparison with the radii  $R_1$  and  $R_2$  of the spherical surfaces and the distance  $f'$  between point  $F'$  and the optical centre of the lens. Therefore, angles  $\gamma_1$ ,  $\gamma_2$  and  $\varphi$  are also small, and we can replace the sines of these angles by the values of these angles (in radians). Further, we can neglect the thickness of the lens assuming that it is thin, and put  $C_1O = R_1$  and  $C_2O = R_2$ . We can also neglect the difference in the heights of points  $M$  and  $M'$  above the axis and assume that they are located at the same height  $h$ . Thus, we can approximately write

$$\gamma_1 \approx \sin \gamma_1 = \frac{h}{R_1}, \quad \gamma_2 \approx \sin \gamma_2 = \frac{h}{R_2}, \quad \varphi \approx \sin \varphi = \frac{h}{f'}. \quad (10.2.6)$$

Substituting these expressions into (10.2.5), we obtain

$$\frac{h}{f'} = (n - 1) \left( \frac{h}{R_1} + \frac{h}{R_2} \right). \quad (10.2.7)$$

Cancelling out  $h$ , we get

$$\frac{1}{f'} = (n - 1) \left( \frac{1}{R_1} + \frac{1}{R_2} \right), \quad (10.2.8)$$

whence

$$f' = \frac{1}{(n - 1) \left( \frac{1}{R_1} + \frac{1}{R_2} \right)}. \quad (10.2.9)$$

It is essential that  $h$  does not appear in the final result. This means that any ray incident on a lens parallel to the principal optical axis at any (but sufficiently small in comparison with  $R_1$  and  $R_2$ ) distance  $h$  from the axis will pass after the refraction in the lens through the same point  $F'$  lying at a distance  $f'$  from the optical centre of the lens.

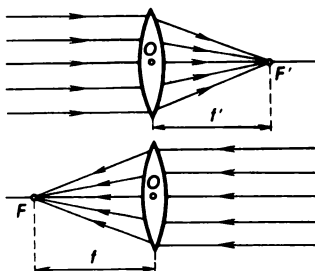


Fig. 196.  
Foci of a lens.

Thus, we have proved that the lens has the principal focus, and formula (10.2.9) shows how the focal length depends on the refractive index of the substance of which the lens is made and on the radii of curvature of its refracting surfaces.

It was assumed that a parallel beam of light is incident on the lens from left to right. The situation will not change if we direct the same beam of rays in the opposite direction, i.e. from right to left. This beam of rays parallel to the principal optical axis will again converge at a single point  $F$ , i.e. the second focus of the lens (Fig. 196) at a distance  $f$  from its optical centre. On the basis of (10.2.9) we conclude that  $f = f'$ , i.e. the two foci of the lens are arranged symmetrically on both its sides.<sup>1</sup>

Focus  $F$  is usually called the *front focus* and  $F'$ , the *rear focus* of the lens. Accordingly, distance  $f$  is known as the *front focal length* and  $f'$ , the *rear focal length* of the lens.

If we place a point light source at a focus of a lens, each ray emanating from this point and refracted by the lens will propagate parallel to the principal optical axis in accordance with the reversibility principle of light rays (see Sec. 9.4). Therefore, a beam of rays parallel to the principal optical axis will emerge from the lens in this case.

When the relations obtained above are used in practice, we should always remember simplifying assumptions introduced in the derivation of these relations. It was assumed that parallel rays are incident on the lens at a very small distance from the axis. This condition is not satisfied quite strictly. Therefore, the points of intersection of refracted rays will not strictly coincide but will occupy a finite region. If we place a screen in this region, a more or less blurred bright spot will be formed on it rather than a geometrical point.

Another circumstance that should be borne in mind is that it is impossible to obtain a strictly point-like light source. Therefore, when we place a source of small but finite dimensions at a focus of a lens, we cannot obtain a strictly parallel beam of rays with the help of the lens.

<sup>1</sup> This conclusion is related to the fact that we assume from the very outset that the same medium (air) is on both sides of the lens. Otherwise, the symmetry in the arrangement of foci  $F$  and  $F'$  would be violated.

It was pointed out in Sec. 8.3 that a strictly parallel beam of rays has no physical meaning. The above remark shows that the properties of the lens considered earlier are in agreement with this general physical statement.

In each individual case when a lens is combined with a certain light source for obtaining a parallel beam of rays or, on the contrary, for focus using a parallel beam, the degree of deviation from the simplifying assumptions under which the formulas have been derived should be specially verified. However, the essential features of refraction of light rays in a lens are correctly determined by these formulas. The deviations from these formulas will be considered later.

### 10.3. Images of Points Located on the Principal Optical Axis of a Lens. Lens Equation

Let a point light source be located at point  $S$  on the principal optical axis of a lens at a distance  $a$  from its optical centre  $O$  (Fig. 197). Let us find out how a narrow beam of rays adjoining the straight line  $SO$  which is the axis of this beam<sup>2</sup> is refracted by the lens.

Let a ray ( $SM$ ) from the light beam be incident on the first (front) refracting surface of the lens at point  $M$  lying at a height  $h$  above the axis. The fact that we limit ourselves to a narrow beam of rays indicates that  $h$  is small in comparison with distance  $a$  from the source to the lens. On the other hand, as in Sec. 10.2, we shall assume that  $h$  is small as compared to  $f'$  and hence as compared to radii  $R_1$  and  $R_2$  of the surfaces bounding the lens. The angle made by ray  $SM$  and the axis will be denoted by  $\gamma$ . Since  $h$  is small, angle  $\gamma$  is small as well. The refracted ray will propagate

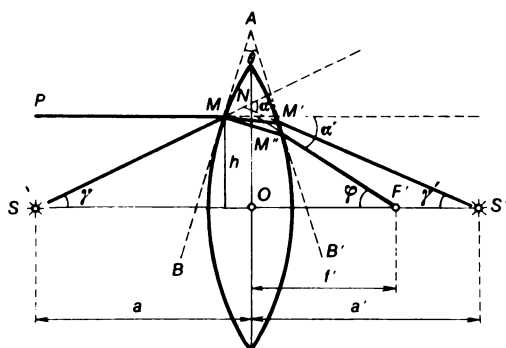


Fig. 197.

Refraction of ray  $SM$  emerging from point  $S$  lying on the principal optical axis of a lens. Angle  $BAB'$  and the lens thickness are strongly exaggerated.

<sup>2</sup> Such beams are usually referred to as *paraxial beams*.

along  $MM'$ , and after the refraction at the second (rear) surface bounding the lens, will emerge from it along  $M'S'$  making angle  $\gamma'$  with the axis. We denote by  $a'$  the distance from the optical centre of the lens to the point  $S'$  at which the refracted ray intersects the principal optical axis.

As in the previous section, we draw through points  $M$  and  $M'$  the planes tangential to the refracting surfaces of the lens. These planes form a thin prism  $BAB'$  with the apical angle  $\theta$ . Instead of analyzing the refraction of ray  $SMM'S'$  in the lens, we shall consider the refraction of this ray in the thin prism  $BAB'$ .

After the refraction, the selected ray will be deflected through angle  $\alpha$  which is determined by the formula for a thin prism:

$$\alpha = (n - 1)\theta, \quad (10.3.1)$$

where  $n$  is the refractive index of the substance of which the lens is made.

Let us also consider ray  $PM$  parallel to the principal optical axis and incident on the lens at point  $M$ . The refraction of this ray has already been considered in Sec. 10.2 (the condition of smallness of  $h$  is observed here). It is well known that after the refraction in the lens, this ray will emerge from point  $M''$  at an angle  $\varphi$  to the axis and will pass through the principal focus  $F'$  at a distance  $f'$  from the optical centre. Points  $M'$  and  $M''$  are very close to each other so that the prisms formed by the tangential planes at point  $M$  and points  $M'$  and  $M''$  are practically indistinguishable and have the same apical angle  $\theta$ . Angle  $\alpha'$  through which this ray will be deflected from its original direction after the refraction in the thin prism is again equal to  $(n - 1)\theta$ , i.e. to angle  $\alpha$ . On the other hand, angle  $\alpha'$  is obviously equal to angle  $\varphi$  (see Fig. 197). Thus, we obtain

$$\alpha' = \alpha = \varphi. \quad (10.3.2)$$

But angle  $\alpha$  is equal to the sum  $\gamma + \gamma'$  as an external angle in triangle  $SNS'$ . Therefore, we have

$$\gamma + \gamma' = \varphi. \quad (10.3.3)$$

Rays  $SM$ ,  $M'S'$  and  $M''F'$  make small angles with the axis, i.e. angles  $\gamma$ ,  $\gamma'$  and  $\varphi$  are small. Replacing, as in the previous section, the sines of the small angles by the values of these angles (in radians) and neglecting the thickness of the lens in comparison with the difference in the heights of points  $M$ ,  $M'$  and  $M''$  above the axis, we can approximately write

$$\gamma \approx \sin \gamma = \frac{h}{a}, \quad \gamma' \approx \sin \gamma' = \frac{h}{a'}, \quad \varphi \approx \sin \varphi = \frac{h}{f'}. \quad (10.3.4)$$

Substituting these approximate expressions into (10.3.3) we obtain

$$\frac{h}{a} + \frac{h}{a'} = \frac{h}{f'}. \quad (10.3.5)$$

Cancelling out the common multiplier  $h$ , we get

$$\frac{1}{a} + \frac{1}{a'} = \frac{1}{f'}. \quad (10.3.6)$$

The quantity  $1/f'$  on the right-hand side of this expression depends, as was shown in the previous section, only on the properties of the lens, i.e. the refractive index of the substance of which the lens is made and on the radii of curvature of the refracting surfaces of the lens.

The fact that formula (10.3.6) does not contain  $h$  leads us to very important conclusions: not only ray  $SM$  but also any other ray emerging from point  $S$  will pass after the refraction in the lens through the same point  $S'$ , although different rays are incident on the lens at different heights above the axis. The only but very important limitation imposed on the rays under consideration consists in that they make small angles with the lens axis.

Thus, all the rays of a narrow beam emerging from point  $S$  will converge after the refraction in the lens again at a single point  $S'$  which is an *image* of point  $S$ . Consequently, we have proved that *the image of a point source lying on the principal optical axis of a thin lens, obtained with the help of a sufficiently narrow beam of rays, is a point.*

Images formed so that each point of an object corresponds to a single point of its image are known as *stigmatic*. Images for which this condition is not observed are called *astigmatic*.<sup>3</sup>

It should be noted that in view of the reversibility principle of light rays (see Sec. 9.4), the positions of a light source  $S$  and its image  $S'$  are reversible. In other words, having placed the source at point  $S'$ , we obtain its image at point  $S$ . Points  $S$  and  $S'$  are termed *conjugate points*.

The problem of obtaining stigmatic images is of special importance in geometrical optics. The stigmatism of images determines the quality of optical systems used for their formation. The violation of stigmatism of light rays incident on an optical system leads to the blurring of the image. Henceforth, while studying simple optical systems, we shall pay much attention to the stigmatism of images formed by them.

Formula (10.3.6) obtained above connects the distances from the optical centre to three points lying on the principal optical axis of a lens: source  $S$ , its image  $S'$  and focus  $F'$ . This is the *basic equation of a thin lens*.

## 10.4. Applications of the Thin Lens Equation.

### Real and Virtual Images

Let us assume that a luminous point  $S$  lying on the principal optical axis of a lens is removed from the lens to a very large distance. In this case,

<sup>3</sup> From the Greek words "stigma" meaning a point and negation "a". Astigmatic means nonpoint-like.

the rays incident on the lens will be nearly parallel to its principal optical axis. It was shown in Sec. 10.2 that after the refraction in the lens, these rays will converge at focus  $F'$  of the lens. As the source is removed to a very large distance, the quantity  $1/a$  in formula (10.3.6) tends to zero, and we get

$$a' = f'.$$

Therefore, we can say that focus  $F'$  is the image of an infinitely remote point.

Any celestial body may serve as an example of an infinitely far body. Consequently, the images of stars, the Sun, etc. will be formed at the focus of a lens. Terrestrial light sources located at far distances from a lens also have images formed at the focus.

Let us now assume that an image of a certain point is formed at a very large distance, i.e. a beam of light rays parallel to the principal optical axis emerges from the lens. It was shown in Sec. 10.2 that in this case the source must be at the front focus  $F$  of the lens (see Fig. 196). This conclusion also follows from formula (10.3.6). Indeed, assuming that the image lies at infinity, we obtain  $1/a' = 0$ . Then the distance from the source to the lens is equal to the focal length:  $a = f = f'$ .

Different lenses are distinguished by the location of the centres of spherical surfaces bounding them, their radii and refractive indices of the substances of which they are made. Figure 198 shows six main types of lenses.

If parallel rays converge after the refraction in a lens so that they actually intersect at a certain point lying on the other side of the lens, the lens is referred to as *converging*, or *positive* (Fig. 199a). If, however, parallel rays become diverging after the refraction in a lens, the lens is *diverging*, or *negative*. In the case of a diverging lens, the imaginary continuations of refracted rays rather than the rays themselves converge at a focus, the

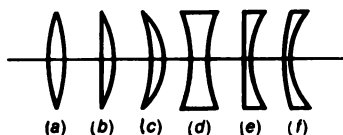


Fig. 198.

Types of lenses. If the lens material refracts more strongly than the surrounding medium, the lenses of types *a*, *b* and *c* are converging, and of types *d*, *e* and *f*, diverging.

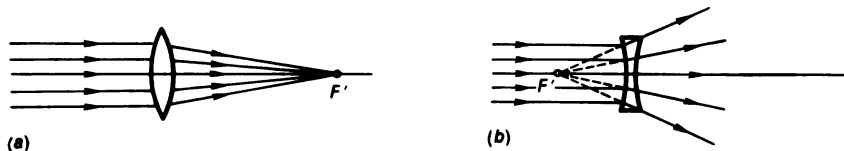
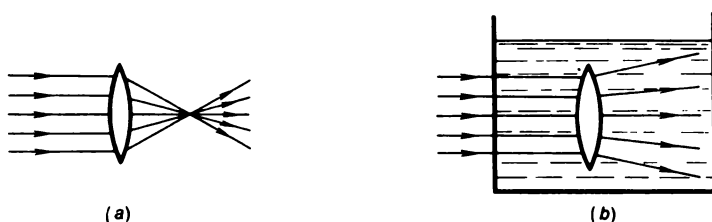


Fig. 199.

The real focus of a converging lens (*a*) and the virtual focus of a diverging lens (*b*).



**Fig. 200.**

Double-convex lenses: (a) a glass lens in air is converging, (b) an air lens in water is diverging.

focus lying on the same side of the lens from which a parallel beam of rays is incident. Foci in such a case are referred to as *virtual* (Fig. 199b).

Normally, the material of a lens refracts stronger than the surrounding medium (e.g. a glass lens in air). Therefore, converging lenses are more thickened in the middle than at the edges (double-convex, plano-convex lenses and a positive meniscus, viz. concavo-convex lens, Figs. 198a-c). Diverging lenses are those which become thinner at the centre (double-concave, plano-concave lenses and a negative meniscus, viz. convexo-concave lens, Figs. 198d-f). If the material of a lens refracts weaker than the surrounding medium, i.e. the relative refractive index  $n < 1$ , then, on the contrary, lenses in Figs. 198a-c are diverging and those in Figs. 198d-f are converging. Such lenses can be obtained, for example, by forming in water an air cavity of appropriate form between two watch glasses (Fig. 200).

Let us now consider luminous points located at a finite distance from a lens. We shall always assume that the sources are to the left of the lens. As to images, image  $S'$  may be to the right or to the left of the lens depending on its type and on the position of the source. If an image is to the right of the lens, this means that it is formed by a converging beam of rays (Fig. 201a), i.e. the rays that actually pass through point  $S'$ . In this case, the image is called *real*. It can be obtained on a screen, photographic plate, and so on. Reconstructing the path of rays forming an image, we can always determine the position of the source, although this involves certain difficulties in actual practice.

**Fig. 201.**

A source and its real image lie on different sides of a lens (a). The virtual image is on the same side of a lens as the source (b).

Let us now assume that an image is to the left of a lens, i.e. on the same side from it as the source. This means that the beam of rays diverging from the source becomes still more divergent after the refraction in the lens, and only imaginary continuations of refracted rays converge at point  $S'$  (Fig. 201b). The image is referred to as *virtual* in this case.

The term “virtual image” which is conventionally used in optics may cause a misunderstanding. In actual practice, there is nothing “virtual” in itself. *The peculiar feature of virtual images is that they cannot be obtained directly on a screen, photographic plate, and so on.* If, for example, we place a very small screen at point  $S'$  (see Fig. 201b), which does not prevent the major part of the rays from getting to the lens, no bright point will be formed on the screen. However, the diverging beam of rays whose imaginary continuations intersect at the virtual image, has nothing “virtual” in itself. This beam can be converted into converging beam if a properly selected converging lens is interposed on its path. Then we shall obtain a real image  $S''$  of the luminous point  $S$  on a screen or a photographic plate (Fig. 202), which can be regarded at the same time as an image of the “virtual point”  $S'$ .

The human eye plays the role of such a converging lens: beams of light diverging from light sources are focussed on the light-sensitive layer, viz. retina. A beam of diverging rays can be focussed to a single point on the retina by the optical system of the eye irrespective of whether these rays emanate from a point source  $S$  or from its virtual image  $S'$ . In everyday life, an observer learns how to automatically reconstruct the path of the rays forming an image on the retina and determine the position of the source. When a diverging beam of rays (with the vertex at point  $S'$ ) shown in Fig. 202 is incident on the eye, we see the source at point  $S'$  by “reconstructing” the point from which these rays emerge, although there is actually no source at this point. This imaginary source is just called a “virtual” source of point  $S$ .

Using formula (10.3.6), we can easily watch how the position of the image changes as the source moves along the principal optical axis (see Ex. II.31 and II.32 at the end of the chapter).



Fig. 202.

Transformation of a diverging beam of rays into a converging beam with the help of an auxiliary converging lens (say, the eye).

### 10.5. Image of a Point Source and of an Extended Object Formed by a Plane Mirror.

#### Image of a Point Source Formed by a Spherical Mirror

Let us now consider the image formation in the light reflected at mirrors of various types. The laws of image formation of luminous points during the reflection at a mirror and the refraction in a lens are similar in many respects.

Naturally, this similarity is not accidental. It is due to the fact that formally (as was shown in Chap. 9), the law of reflection is a particular case of the law of refraction (for  $n = -1$ ).

The problem formulated above is solved most easily for the reflection of light at a plane mirror. At the same time, the reflection of light at a plane mirror is the simplest and the best known case of virtual images considered in the previous section.

Let a beam of rays from a point source  $S$  (Fig. 203) be incident on a plane mirror (metal mirror, water surface, etc.). Let us trace this cone of rays with the vertex at point  $S$ . We take two arbitrary rays  $SA$  and  $SB$ . Each of them will be reflected in accordance with the law of reflection so that the angle between each ray and the normal remains unchanged after the reflection. Consequently, the angle between the two rays also remains the same after the reflection.

This angle between the reflected rays can be constructed on the figure by continuing the reflected rays in the backward direction behind the plane of the mirror, as shown by dashed lines in the figure. The point  $S'$  of intersection of continuation of rays behind the mirror will lie on the same normal to the mirror as point  $S$  and at the same distance from the plane of the mirror. This can easily be verified from the equality of triangles  $SAO$  and  $S'AO$  or  $SBO$  and  $S'BO$ .

Since the rays  $SA$  and  $SB$  under consideration were chosen arbitrarily, we can generalize the results obtained for them and concerning the reflection at a plane mirror to the entire light beam. Consequently, we can state that a beam of light rays emanating from a single point is transformed, as a result of the reflection at a plane mirror, into a light beam in which the continuations of all light rays again intersect at a single point.

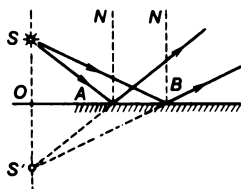
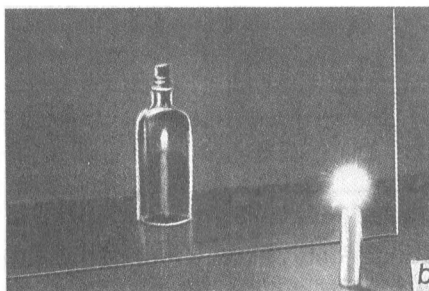
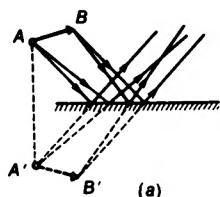


Fig. 203.

Formation of a virtual image of a point by a plane mirror.

**Fig. 204.**

(a) Formation of a virtual image of a straight segment by a plane mirror. (b) It seems to an observer that a candle glows in a bottle with water placed behind a glass plate in the region where the virtual image of the candle is formed by this plate.

As a result, an observer located on the path of reflected rays will see them as intersecting at point  $S'$ , this point being a virtual image of point  $S$ . The image will be virtual in the sense clarified above: there are no rays at point  $S'$  behind the mirror, but this point is the vertex of the light beam reversed as a result of reflection at the mirror.

An analysis of the virtual image of a luminous point formed by a plane mirror and conclusions concerning the position of this image “behind the mirror” make it possible to easily find the image of an extended object formed by a plane mirror.

Let a straight luminous segment  $AB$  be located in front of a mirror (Fig. 204a). Constructing points  $A'$  and  $B'$  as described above and connecting them by the straight line, we obtain the image of all points of the segment. This follows from elementary geometrical considerations. Since segment  $AB$  was chosen arbitrarily, we can construct in this way the image of any object. The normals to the mirror are parallel to one another; hence it is clear that the size of the virtual image formed by a plane mirror is equal to the size of an object placed in front of the mirror.

It should be emphasized for the solution obtained for the reflection of light beams at a plane mirror that the image of each point of a luminous body is a single point (stigmatic image).

Let us now consider spherical mirrors. Figure 205 shows cross section  $APB$  of a concave spherical mirror of radius  $R$ ,  $C$  being the centre of the sphere. The central point of the available part of the spherical surface is known as the pole  $P$  of the mirror. The normal to the mirror passing through its centre and pole is called the *principal optical axis* of the mirror. The normals to the mirror drawn through other points of its surface and naturally passing through the centre  $C$  of the mirror as well are known

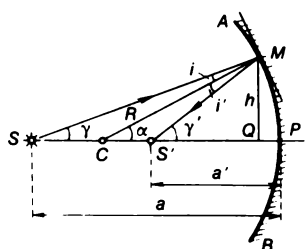


Fig. 205.

Reflection of ray  $SM$  emerging from point  $S$  on the axis by a spherical mirror.

as auxiliary optical axes. One of them ( $MC$ ) is shown in Fig. 205. All normals to the spherical surface are of course equivalent, and the distinction between the principal optical axis and auxiliary axes is immaterial.<sup>4</sup> The diameter of the circle bounding a spherical mirror is known as its *aperture*.

The following digression is a simplified version of what was said in Secs. 10.2 and 10.3 about lenses.

Let a point source  $S$  be located on the principal optical axis of a mirror at a distance  $SP = a$  from the pole. As in the case of lenses, we consider ray  $SM$  belonging to a narrow beam, i.e. making a small angle  $\gamma$  with the axis and incident on the mirror at point  $M$  located at a height  $h$  above the axis so that  $h$  is small in comparison with  $a$  and with the radius  $R$  of the mirror. The reflected ray will intersect the axis at point  $S'$  at a distance  $S'P = a'$  from the pole. The angle made by the reflected ray and the axis will be denoted by  $\gamma'$ . This angle is also small.

Obviously,  $CM$  is a normal to the surface of the mirror at the point of incidence,  $i$  is the angle of incidence and  $i'$  is the angle of reflection. According to the law of reflection, we have

$$i = i'. \quad (10.5.1)$$

We denote by  $\alpha$  the angle made by radius  $CM$  and the axis. From triangle  $SMC$ , we have

$$i + \gamma = \alpha, \quad (10.5.2)$$

and from triangle  $CMS'$ ,

$$\gamma' = \alpha + i'. \quad (10.5.3)$$

Summing up (10.5.2) and (10.5.3) and considering that  $i = i'$ , we obtain

$$\gamma + \gamma' = 2\alpha. \quad (10.5.4)$$

Since we consider a narrow beam of light adjoining the principal optical axis, i.e. angles  $\gamma$ ,  $\gamma'$  and  $\alpha$  are small, we can replace the sines of these angles by the angles themselves and neglect the length of segment  $PQ$ . Then

<sup>4</sup> In thin lenses, the principal optical axis differs considerably from auxiliary axes in that it is the single axis passing through the centres of both spherical surfaces bounding the lens.

we obtain the following approximate equalities:

$$\gamma \approx \sin \gamma = \frac{h}{a}, \quad \gamma' \approx \sin \gamma' = \frac{h}{a'}, \quad \alpha \approx \sin \alpha = \frac{h}{R}. \quad (10.5.5)$$

Substituting these expressions into (10.5.4) and cancelling out the common factor  $h$ , we obtain

$$\frac{1}{a} + \frac{1}{a'} = \frac{2}{R}. \quad (10.5.6)$$

The fact that neither height  $h$  nor angle  $\gamma$  appear in the final result indicates that any ray emanating from point  $S$  (and belonging to a sufficiently narrow beam) will pass after the reflection through point  $S'$  lying at a distance  $a'$  from the pole. Therefore, point  $S'$  is the image of point  $S$ .

Thus, the image of a point source formed by a spherical mirror is also a point. As in the case of a lens, point  $S$  at which the source is located and point  $S'$  corresponding to the image of this point are conjugate. This means that by placing the source at point  $S'$ , we obtain the image at point  $S$  (in accordance with the reversibility principle of light rays, see Sec. 9.4).

Formula (10.5.6) obtained above is the *basic equation of a spherical mirror*. It can easily be proved that this formula is also valid for a convex spherical mirror.

### 10.6. Focal Point and Focal Length of a Spherical Mirror

Let us determine the position  $F$  of the focus of a spherical mirror, i.e. the point at which the rays parallel to the principal optical axis intersect after the reflection at such a mirror. It is well known that in order to obtain a parallel beam of rays, the source must be placed at a long distance, i.e. we must put in formula (10.5.6)  $1/a = 0$ . In this case,  $a' = f$  is the *focal length* of the mirror. Using (10.5.6), we obtain the following expression for the focal length:

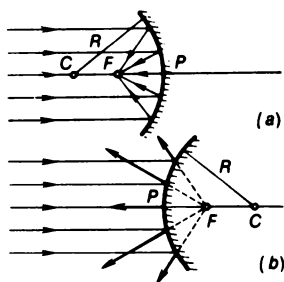
$$f = \frac{R}{2}. \quad (10.6.1)$$

Combining (10.5.6) and (10.6.1), we obtain the equation for a spherical mirror in the form

$$\frac{1}{a} + \frac{1}{a'} = \frac{1}{f}, \quad (10.6.2)$$

i.e. in the form similar to formula (10.3.6) for a thin lens.

For a *concave* mirror, the focus is located at the midpoint of the distance between the pole and the centre to the left of the pole (Fig. 206a). For a *convex* mirror, the focus is located at a distance  $R/2$  to the right of the pole, i.e. is *virtual* (Fig. 206b).

**Fig. 206.**

Foci of spherical mirrors: (a) concave mirror, (b) convex mirror. (The rays are shown to be incident on considerable parts of the mirrors. They should intersect the mirror at a small height from the axis, i.e. embrace a small part of the mirrors.)

Using the fact that the source and its image are at conjugate points, we can immediately draw the following conclusion: if a point source is at the focus of a mirror, its image is at infinity, i.e. a parallel beam of rays emerges from the mirror. This condition forms the basis for obtaining parallel light beams with the help of concave mirrors (to be more precise, nearly parallel beams). As applied to floodlights, this property was considered in Chap. 8.<sup>5</sup>

It should be noted that while considering the properties of a spherical mirror, we assumed, like in the case of a lens, that, firstly, a very narrow beam of rays adjoining the mirror axis is used, and, secondly, a point source is taken. Of course, these requirements cannot be satisfied strictly. The question of whether the deviations from these requirements are essential should be answered while solving a specific problem.

### 10.7. Relation Between the Positions of a Source and Its Image on the Principal Optical Axis of a Spherical Mirror

Let us trace the change in the position of the image of a light source approaching a concave mirror from infinity (Figs. 207*a-d*). It follows from formula (10.6.2) that if the source moves from infinity to the centre of the mirror, its image is displaced from the focus to the centre of the mirror. Ultimately, the positions of the source and its image coincide (Fig. 207*b*).

As the source moves from the centre to the focus, its image moves away from the centre of the mirror (Fig. 207*c*). When the light source is located at the focus, its image goes to infinity. In other words, if a point source is located at the focus of the mirror, it produces a parallel beam of rays.

Finally, if the light source is between the focus of the mirror and its pole, the reflected rays will not have a common vertex on the concave side of the mirror and will not intersect the principal optical axis of the mirror

<sup>5</sup> In Chap. 8, the mirror of a floodlight had the paraboloidal rather than the spherical shape. This mirror gives a nearly parallel beam of light even for a considerable size of the aperture, while a spherical mirror satisfies this conditions only for small apertures (small values of  $h$ ).

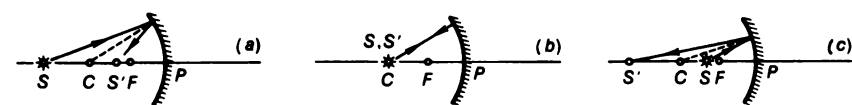


Fig. 207.

Image formation by a concave spherical mirror for different positions of a point source on the mirror axis: (a) the source is between the centre of the mirror and infinity, (b) the source is at the centre, (c) between the centre and the focus, (d) between the focus and the mirror.

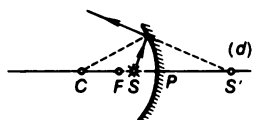
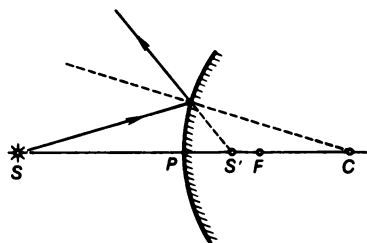


Fig. 208.

Image construction for a convex spherical mirror.



(Fig. 207d). Only their continuations behind the mirror (shown by the dashed lines in the figure) will have the vertex in common ( $S'$ ). This means that the image is virtual in this case. In order to determine the position of the image, it is sufficient to extend the direction of any reflected ray behind the mirror. The point of its intersection with the principal optical axis will give the position of the image.

Let us consider images formed by spherical convex mirrors. It was mentioned above that a convex mirror has a virtual focus at a distance  $R/2$  from the pole. The construction of the image of a point located at a finite distance from the mirror is shown in Fig. 208. It can be seen that the image produced by a convex mirror is always virtual.

### 10.8. Methods of Preparation of Lenses and Mirrors

*Optical glass* is the basic material used for preparing lenses, prisms and other optical elements. Glass is transparent and can be extremely homogeneous. It is important that glass is stable to mechanical as well as chemical effects. For this reason, glass parts can be finished to a very high degree of accuracy, and the shape given to them thereby remains unchanged.

The optical properties of glass (above all, its refractive index) can be varied over a wide range by an appropriate change in its composition. The main part of glass is silicon dioxide  $\text{SiO}_2$ , to which oxides of other elements (sodium, potassium, calcium, barium, aluminum, boron, lead, etc.) are added. Depending on the type and amount of admixtures, the optical properties of glass vary.

Optical glass intended for preparing an optical element is first cut and roughly processed to give it the required shape and size. Then it is subjected to grinding and polishing. As a rule, the treatment of optical parts is carried out to a very high degree of accuracy (the deviation from a given curvature must be within  $0.00002$  mm). The accuracy requirements here are 500 times higher than for ordinary processing of mechanical parts, carried out by mechani-



cal measuring instruments. For this reason, the quality of finishing is normally controlled by using special optical methods based on interference.

In mirrors used for household, the reflecting layer is deposited on the back surface of a glass plate and can be seen only through the glass. The layer is deposited by chemical methods involving the precipitation of metal silver from  $\text{AgNO}_3$  solution with admixtures of some substances. Such a layer is usually protected from the rear surface by a layer of lacquer, covered with a sheet of cardboard or wood, and its front surface is covered with the glass, which prolongs the service life of mirrors.

This method, however, is inapplicable for mirrors used in scientific research laboratories since a mirror obtained by the method described above produces a weak additional reflection (about 5%) from the front surface of the glass, and the rays reflected from the metal layer must pass through the glass layer, which slightly changes their direction and considerably complicates calculations. For this reason, the well-reflecting metal layer is deposited meticulously on the ground and polished front surface of the glass. A layer of silver or aluminum used for this purpose is deposited by vacuum evaporation or cathode sputtering. Fresh layers of such metals have reflection coefficients up to 90% and even higher. Unfortunately, the reflectivity of mirrors with “outer” coating becomes worse with time. Recently, very stable mirrors with reflection coefficients of up to 95% and higher have been obtained by coating glass with several layers of various (non-metallic) materials of strictly selected thickness. The high reflectivity of such multilayer coatings is due to interference of light.

## 10.9. Images of Extended Objects

### Formed by Spherical Mirrors and Lenses

We assumed so far that a light source is a luminous point located at the principal optical axis of a mirror or lens. Let us now consider the images of small objects arranged near the principal optical axis of a spherical mirror or lens. The expression “small object” will mean that a given object is seen from the centre of a mirror or lens at a small angle. Since individual points of an extended object lie outside the principal optical axis, the problem is reduced to the construction of images of such “off-axis” points. This problem can easily be solved. Let us analyze it for the case of a spherical mirror.

Let a point source  $S_1$  be at a certain distance from the principal optical axis of the mirror (Fig. 209). We draw through it an auxiliary optical axis. As regards the reflection at the spherical mirror, point  $S_1$  is equivalent to point  $S$  lying on the principal axis of the mirror at the same distance from its centre  $C$ . Thus, using the results of Sec. 10.5, for a narrow beam of rays isolated around the axis  $S_1C$ , we can state that after the reflection it will be focussed again at a single point  $S'_1$  which is the image of point

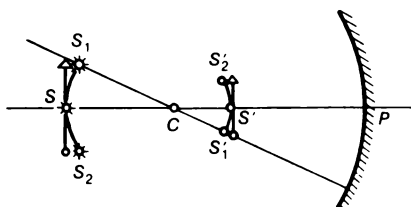


Fig. 209.

Construction of the image of an extended object for a spherical mirror.

$S_1$ . It can easily be seen that the image of any point belonging to arc  $S_1SS_2$  with the centre at point  $C$  will be a point belonging to arc  $S_1'S'S_2'$  with the centre at the same point  $C$ . In other words, arc  $S_1'S'S_2'$  is the image of arc  $S_1SS_2$ .

We shall assume that all points of arc  $S_1SS_2$  are at a small distance from the principal optical axis of the mirror. Then we can replace arcs  $S_1SS_2$  and  $S_1'S'S_2'$  by rectilinear segments normal to the principal optical axis.

Thus, we have proved that *the image of a small segment perpendicular to the principal axis of a spherical mirror will also be a segment perpendicular to the principal optical axis*. This conclusion is valid only if the angle at which the object is seen from the centre of the mirror is sufficiently small. Otherwise, the arc cannot be replaced by a rectilinear segment. The violation of this condition in actual practice makes the image unclear and blurred at the edges.

The problem for a thin lens is solved in a similar way. In this case too, a clear image of an extended object is obtained only provided that this object (its extreme points) is seen from the optical centre of the lens at a small angle to the principal optical axis. If this condition is violated, the image is more or less blurred and distorted.

## 10.10. Magnification of Images

### Formed by Spherical Mirrors and Lenses

Let us now consider the size of the image formed by a mirror or lens. The constructions shown in Fig. 210 indicate that in contrast to the plane mirror, the size of the image formed by the spherical mirror will change depending on the position of an object relative to the focus of the mirror. For example, if the object is far from the focus of a concave mirror, its image will be diminished. If the object is between the mirror and its focus, the image will be virtual and magnified.

The ratio of linear sizes of image  $S_1'S_2' = y'$  and  $S_1S_2 = y$  of an object is called the *linear, or transverse, magnification*:

$$\beta = \frac{y'}{y} = \frac{S_1'S_2'}{S_1S_2}.$$

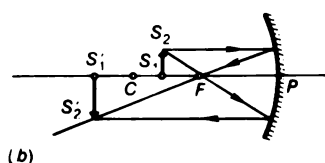
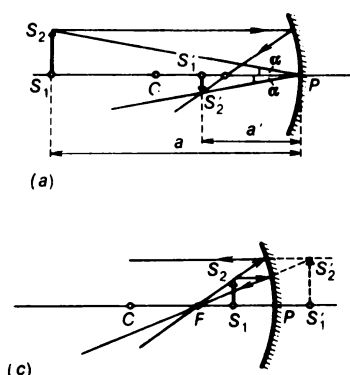


Fig. 210.

Formation of images of extended objects by a concave spherical mirror. An object is placed (a) behind the centre of the mirror (the image is real, inverted and reduced), (b) between the centre and the focus (the image is real, inverted and magnified), (c) in front of the focus (the image is virtual, erect and magnified).

From the similarity of triangles  $S_1PS_2$  and  $S'_1PS'_2$  (Fig. 210a), we obtain

$$\beta = \frac{y'}{y} = \frac{a'}{a}. \quad (10.10.1)$$

It can easily be seen that this equality is also valid for other cases of image formation with the help of spherical mirrors (Figs. 210b and c).

The images formed by a lens can also be either magnified or diminished. From the similarity of triangles  $S_1OS_2$  and  $S'_1OS'_2$  (Fig. 211), we get the same expression for the magnification of the lens as that obtained earlier for a spherical mirror:

$$\beta = \frac{y'}{y} = \frac{a'}{a}. \quad (10.10.2)$$

Besides the linear magnification, we shall also consider the *angular magnification* of a lens (or spherical mirror). The angular magnification  $\gamma$  is the ratio of the tangents of angles  $\alpha'$  and  $\alpha$  formed by the ray emerging from the lens and the incident ray with the principal optical axis:

$$\gamma = \frac{\tan \alpha'}{\tan \alpha}. \quad (10.10.3)$$

Figure 212 shows that

$$h = a \tan \alpha = a' \tan \alpha',$$

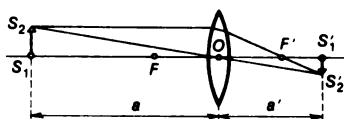


Fig. 211.

Linear magnification of a lens,  $\beta = \frac{S'_1S'_2}{S_1S_2} = a'/a$ .

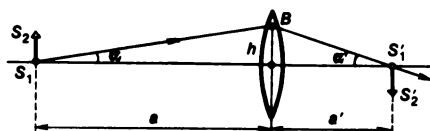


Fig. 212.

Angular magnification of a lens,  $\gamma = \frac{\tan \alpha'}{\tan \alpha} = a/a'$ .

whence

$$\gamma = \frac{\tan \alpha'}{\tan \alpha} = \frac{a}{a'}.$$

Comparing this relation with (10.10.1), we find that

$$\gamma = \frac{1}{\beta}, \quad (10.10.4)$$

i.e. *the angular magnification is reciprocal to the linear magnification*. Hence it follows that the larger the linear magnification, i.e. the size of the image, the smaller the angular magnification, i.e. the narrower the beams of light rays forming the image. This circumstance plays an important role for determining the brightness of the image (see Chap. 11).

### 10.11. Image Formation by Spherical Mirrors and Lenses

Constructing the image of any point of a source, there is no need to consider a large number of rays. It is sufficient to construct two rays, and the point of their intersection will determine the location of the image. It is most convenient to choose the rays whose paths can easily be traced. The paths of these rays in the case of the reflection at a spherical mirror are shown in Fig. 213.

Ray 1 passes through the centre on the mirror and hence is normal to its surface. This ray is reflected back exactly along the principal or an auxiliary optical axis.

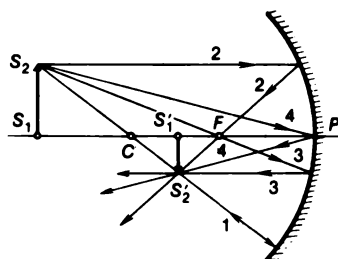
Ray 2 is parallel to the principal optical axis of the mirror. After the reflection, it will pass through the focus of the mirror.

Ray 3 passes from the point of an object through the focus of the mirror. After the reflection, it will propagate parallel to the principal optical axis of the mirror.

Ray 4 incident on the mirror at its pole is reflected back symmetrically about the principal optical axis.

*For constructing an image, we can use any pair of these rays.*

Having constructed the images of a sufficient number of points of an extended object, we can get an idea about the position of the image of the entire object. For objects having a simple shape like that shown in



**Fig. 213.**

Various methods of image construction by a concave spherical mirror.

Fig. 213 (a segment of a straight line perpendicular to the principal optical axis), it is sufficient to construct just one point  $S'_2$  of the image. More complicated cases are considered in exercises.

Figure 210 represents geometrical constructions of the images for different positions of an object in front of the mirror. Figure 210c, where the object is placed between the mirror and the focus, illustrates the construction of the virtual image with the help of continuations of rays behind the mirror.

Figure 214 shows the construction of the image in the case of a convex mirror. As was mentioned above, here we always obtain virtual images.

In order to construct the image of any point of an object formed by a lens, it is sufficient, like in the construction of the image formed by a mirror, to find the point of intersection of any two rays emanating from this point. Figure 215 shows the rays with which the construction can easily be carried out.

Ray 1 passes along an auxiliary optical axis without changing its direction.

Ray 2 is incident on the lens parallel to the principal optical axis. After the refraction, it will pass through the rear focus  $F'$ .

Ray 3 passes through the front focus  $F$ . After the refraction, it will emerge parallel to the principal optical axis.

These rays are constructed without any difficulty. Any other ray emerging from point  $S_2$  would be more difficult to construct since the law of refraction had to be used directly. But this is not necessary since any other refracted ray will pass through the same point  $S'_2$ .

It should be noted that while constructing the images of "off-axis" points, it is not necessary that all selected pairs of rays actually pass through the lens (or mirror). In many cases, like in photography, an object is considerably larger than a lens, and rays 2 and 3 (Fig. 216) do not pass through the lens. Nevertheless, these rays can be used for constructing the image. Actual rays participating in the image formation are bounded by the lens mount (hatched cones), but they naturally converge at the same point  $S'_2$  since it was proved that the image of a point source formed by a lens is a point.

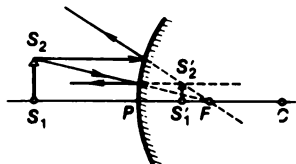


Fig. 214.

Image construction by a convex spherical mirror.

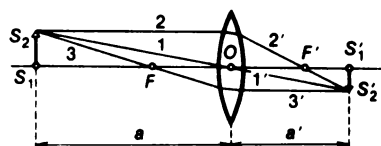


Fig. 215.

Various methods of image construction by a lens.

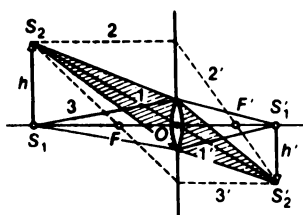


Fig. 216.

Image construction in the case when an object is considerably larger than a lens.

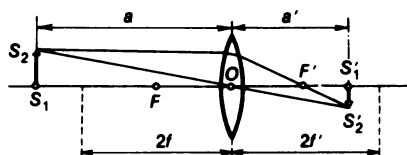


Fig. 217.

Image construction by a lens when an object is behind the double focal length.

Let us consider a few typical cases of image formation by a lens. In cases 1-5 we shall assume that the lens is *converging*.

1. *An object is separated from the lens by a distance exceeding the double focal length.* Such is the normal position of an object to be photographed.

The construction of the image is shown in Fig. 217. Since  $a > 2f$  according to the lens equation (10.3.6) we have

$$\frac{1}{a'} = \frac{1}{f} - \frac{1}{a} > \frac{1}{2f}, \quad a' < 2f,$$

i.e. *the image is between the rear focus and the point lying at a double focal length from the optical centre of the lens. The image is inverted and diminished* since the magnification

$$\beta = \frac{a'}{a} < 1.$$

2. Let us consider an important special case when *a beam of rays parallel to an auxiliary optical axis is incident on a lens*. This is observed in photographing very remote extended objects.

The construction of the image is shown in Fig. 218. In this case, *the image lies on the corresponding auxiliary optical axis at the point of its intersection with rear focal plane* (this is the term applied to the plane perpendicular to the principal optical axis and passing through the rear focus of the lens).

The points of the focal plane are often referred to as the *foci* of the corresponding auxiliary optical axes, point  $F'$  being called the *principal focus* corresponding to the principal optical axis.

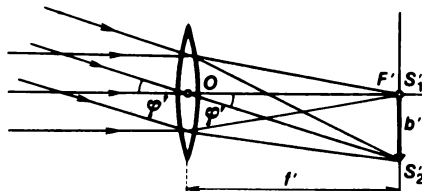


Fig. 218.

Image construction in the case when a beam of rays parallel to an auxiliary optical axis is incident on a lens.

Distance  $b'$  from focus  $S_2'$  to the principal optical axis and angle  $\varphi'$  between the auxiliary optical axis under consideration and the principal optical axis are obviously related through the formula (see Fig. 218)

$$\tan \varphi' = \frac{b'}{f'}. \quad (10.11.1)$$

3. *An object is between the point at a double focal length and the front focus.* This is the normal position of the object projected by a projection lantern.

In order to analyze this case, it is sufficient to make use of the reversibility of image formed by a lens. We shall assume that  $S_1'S_2'$  is the source (see Fig. 217); then  $S_1S_2$  will be the image. It can easily be seen that in the case under consideration, *the image is inverted and magnified and lies at a distance exceeding the double focal length.*

It is useful to consider a particular case when *an object is at a distance equal to the double focal length ( $a = 2f$ )*. Then according to the lens equation, we have

$$\frac{1}{a'} = \frac{1}{f} - \frac{1}{2f} = \frac{1}{2f}, \quad a' = 2f,$$

i.e. *the image also lies at the double focal length from the lens. The image is inverted.* The magnification is

$$\beta = 1,$$

i.e. *the size of the image is the same as that of the object.*

4. Of great importance is the particular case when *the source is in the plane perpendicular to the principal optical axis of the lens and passing through the front focus.*

This plane is also a focal plane referred to as the *front focal plane*. If a point source is at a point of the focal plane, i.e. at one of the front foci, a parallel beam of rays directed along the corresponding optical axis emerges from the lens (Fig. 219). Angle  $\varphi$  between this axis and the principal optical axis and distance  $b$  from the source to this axis are connected through the relation

$$\tan \varphi = \frac{b}{f}. \quad (10.11.2)$$

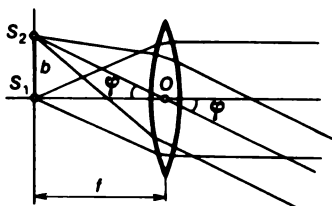


Fig. 219.

Sources  $S_1$  and  $S_2$  lie in the front focal plane. (Beams of rays emerging from the lens are parallel to the auxiliary optical axes passing through the sources.)

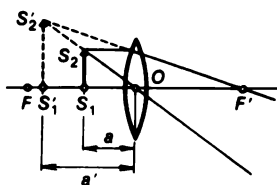


Fig. 220.

Image construction for the case when an object is between the front focus and a lens.

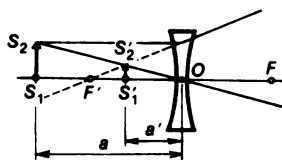


Fig. 221.

Image construction for a diverging lens.

5. An object is between the front focus and the lens, i.e.  $a < f$ . In this case, the image is erect and virtual.

The construction of the image in this case is shown in Fig. 220. Since  $a < a'$ , for the magnification we have

$$\beta > 1,$$

i.e. the image is magnified. We shall return to this case while considering a magnifying glass.

6. Image construction for a diverging lens (Fig. 221).

The image formed by a diverging lens is always virtual and erect. Finally, since  $a' < a$ , the image is always diminished.

It should be noted that in all cases of construction of rays passing through a *thin* lens, we can neglect the path of the rays within the lens itself. It is only important to know the positions of the optical centre and of the principal foci. Thus, a thin lens can be represented by a plane passing through the optical centre of the lens perpendicular to the principal optical axis on which the positions of the principal foci are marked. This plane is known as the *principal plane*. Obviously, the ray entering the lens and emerging from it passes through the *same* point on the principal plane (Fig. 222a). If in the figures we preserve the contours of lenses, this is done only

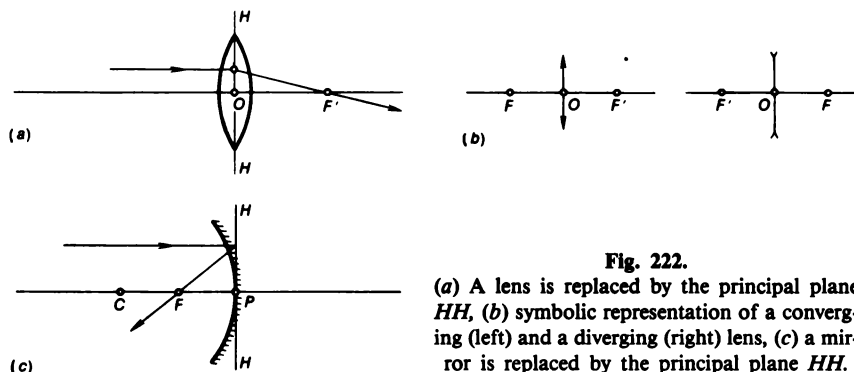


Fig. 222.

(a) A lens is replaced by the principal plane  $HH$ , (b) symbolic representation of a converging (left) and a diverging (right) lens, (c) a mirror is replaced by the principal plane  $HH$ .



for visual distinction between converging and diverging lenses; these contours are superfluous for constructions. Sometimes, in order to simplify figures, the lenses are symbolically represented as in Fig. 222*b*.

Similarly, a spherical mirror can be represented by the principal plane tangential to the surface of the sphere at the mirror pole, the positions of the centre  $C$  of the sphere and the principal focus  $F$  being indicated on the principal optical axis. The position of  $C$  indicates whether we have a concave (converging) or convex (diverging) mirror (Fig. 222*c*).

### 10.12. Optical Power of Lenses

The optical properties of various lenses are often characterized by a quantity reciprocal to the focal length  $f$  of the lens. The quantity

$$D = \frac{1}{f} \quad (10.12.1)$$

is known as the *optical power* of the lens.

The shorter the focal length, the stronger the refraction in the lens and the larger the value of  $D$ . Thus,  $D$  may serve as the characteristic of the *refractivity* of the lens.

The unit of optical power is the optical power of a lens with the focal length of 1 m. This unit is known as a *dioptr* (D). The optical power of any lens (in dioptr) is equal to unity divided by the focal length (in metres). For converging (positive) lenses, the optical power is positive, while for diverging (negative) lenses, it is negative. For example, for a diverging lens with a focal length  $f = 20$  cm, the optical power is  $D = -1/0.20 = -5$  D.

This notation is well known for those who wear glasses.

? II.28. Using the method employed for deriving the lens equation, derive the formula for the refraction at the spherical interface between two media (say, air and glass, Fig. 223).

II.29. Prove that the focal length of the spherical surface mentioned in Ex. II.28 are connected through the relation

$$\frac{f_1}{f_2} = \frac{n_1}{n_2},$$

where  $n_1$  and  $n_2$  are the refractive indices of the first and the second medium.

II.30. Determine the focal length of a plano-convex lens for which the radius of curvature of the spherical surface is 80 cm. The refractive index of glass is 1.6.

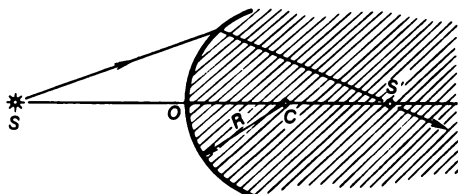


Fig. 223.  
To Exercise II.28.

**II.31.** A converging lens has a focal length of 40 cm. An object is 1 m from the lens. Determine the position of the image and the linear and angular magnifications. Solve the problem analytically and graphically by plotting (to a certain scale) the image of a small object formed by the lens.

**II.32.** Solve the previous problem for the case when an object is at a distance of 20 cm from the lens.

**II.33.** A concave mirror has a radius of 40 cm. An object is at a distance of 30 cm from the mirror. Determine the position of the image and the magnification of the mirror. Plot the image of the object and determine whether it is erect or inverted.

**II.34.** Indicate the position of the image formed by a thin lens if a source is on the principal optical axis (a) at infinity; (b) at a double focal length; (c) at the principal focus, and (d) between the principal focus and the lens.

**II.35.** Analyze the change in the position and size of the image as a result of the change in the position of an object for (a) converging lens; (b) diverging lens; (c) concave mirror, and (d) convex mirror.

**II.36.** The optical power of a lens is 2 D. Determine its focal length.

**II.37.** Plot the image of a small segment formed by a lens if the segment is tilted to the axis at  $45^\circ$ .

**II.38.** A plane mirror is rotated through an angle  $\beta$  about an axis lying in the plane of the mirror and perpendicular to the incident ray. By what angle will the reflected ray be turned thereby?

## Chapter 11

# Optical Systems and Errors

### 11.1. Optical System

A thin lens is the *simplest optical system*. Simple thin lenses are mainly used in glasses. Besides, lenses are used as magnifying glasses (magnifiers).

As was mentioned in Sec. 10.11, the operation of many optical instruments (projection lantern, photographic camera, etc.) can be represented schematically as the operation of thin lenses. A thin lens, however, produces a good image only in comparatively rare cases when we can limit ourselves to a narrow monochromatic beam propagating from a source along the principal axis or at a small angle to it. In most practical problems where these conditions are violated, the image produced by a thin lens is imperfect. For this reason, more *complex optical systems* are designed, which contain a large number of refracting surfaces and are not subjected to the requirement that these surfaces should be close (the requirement satisfied for a thin lens).

### 11.2. Principal Planes and Principal Points of a System

Let us construct a complex optical system by arranging a few lenses one behind another so that their principal axes coincide (Fig. 224). This principal axis, which is the same for the entire system, passes through the centres of all surfaces bounding individual lenses. We direct a beam of parallel rays on the system so that the condition of smallness is observed for the diameter of this beam, as in Sec. 10.2. It will be seen that behind this system, the beam converges at a single point  $F'$  which will be called the *rear focus of the system* as in the case of a thin lens. Having directed a parallel beam on the system from the opposite side, we determine the *front focus of the system* ( $F$ ). It is difficult, however, to determine the focal length of the system since we cannot say to which point of the system this

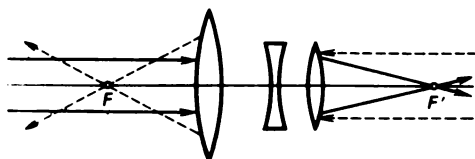


Fig. 224.

Foci of an optical system.

distance should be measured from points  $F$  and  $F'$ . Generally, in the optical system there is no point similar to the optical centre of a thin lens, and we have no grounds to give preference to any of the surfaces comprising the system. In particular, the distances from  $F$  and  $F'$  to the corresponding outer surfaces of the system are not equal.

These difficulties are removed as follows.

For a thin lens, all constructions can be made without analyzing the path of the rays in the lens, and we can confine ourselves to representing the lens by its principal plane (see Sec. 11.11).

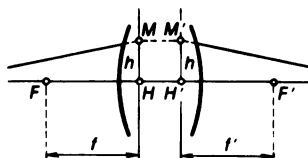
An analysis of the properties of complex optical systems shows that in this case too we can do without considering the actual path of the rays in the system. However, a complex optical system can be replaced not by one but by a set of *two principal planes* perpendicular to the optical axis of the system and intersecting the axis at two so-called *principal points* ( $H$  and  $H'$ ). Having marked the positions of the principal foci on the axis, we shall have a complete characteristic of the optical system (Fig. 225). Now the contours of the outer surfaces bounding the system (shown by solid arcs in Fig. 225) are superfluous. Two principal planes of the system replace a single principal plane of a thin lens: a transition from a system to a thin lens indicates that the two principal planes come closer to each other till they merge into one so that the principal points  $H$  and  $H'$  come closer and ultimately coincide with the optical centre of the lens.

Therefore, the principal planes of the system as if correspond to the decomposition of the principal plane of a thin lens. This circumstance is in agreement with the basic property of the principal planes: a ray entering the system intersects the first principal plane at the *same height*  $h$  as that at which the emerging ray intersects the second principal plane (see Fig. 225).

We shall not prove the fact that such a pair of planes actually exists in any optical system, although such a proof is fairly simple. We shall limit ourselves only to finding the method of employment of these characteristics for constructing images.

Principal planes and principal points may lie either inside or outside the system asymmetrically to the surfaces bounding the system (for example, on the same side of it).

Using principal planes, we can determine the focal lengths of the system. *Focal lengths of an optical system are the distances from the principal*



**Fig. 225.**  
Principal planes of an optical system.

points to the corresponding foci. Therefore, if we denote by  $F$  and  $H$  the front focus and the front principal point and by  $F'$  and  $H'$  the rear focus and the rear principal point, then  $f' = H'F'$  is the rear focal length of the system, while  $f = HF$  is its front focal length.

If the same medium (say, air) is on both sides of the system so that the front and rear foci lie in this medium we have

$$f = f' \quad (11.2.1)$$

as for a thin lens.

### 11.3. Image Construction in a System

If we know the position of the principal and focal planes of a system, we can construct the image without going into details concerning its specific properties (the number of refracting surfaces, their position, curvature, and so on). For image construction, it is sufficient to draw any two rays from those which can easily be constructed. The path of these rays is shown in Fig. 226.

Ray 1 is incident on the system parallel to the principal axis. If this ray intersects the front principal plane at point  $Q$ , then according to the property of principal planes, it will intersect the rear principal plane at point  $Q'$  at the same height above the axis and, having passed through the system, will cross the principal axis at point  $F'$ .

Ray 2 passes through the front focus and intersects the principal axis at a point  $R$ . It will pass through the rear principal plane at the same height ( $R'H' = RH$ ) and will emerge from the system parallel to the principal axis.

This pair of rays can be used for constructing the image of point  $S_2$  in the given system. Accordingly, the image of segment  $S_1S_2$  will be segment  $S'_1S'_2$ .

### 11.4. Magnification of a System

Let us now derive the formulas for the *linear magnification*  $\beta$  of a system. From the similarity of triangles  $S'_1S'_2F'$  and  $H'Q'F'$  (Fig. 226) we have

$$\frac{S'_1S'_2}{H'Q'} = \frac{F'S'_1}{F'H'}.$$

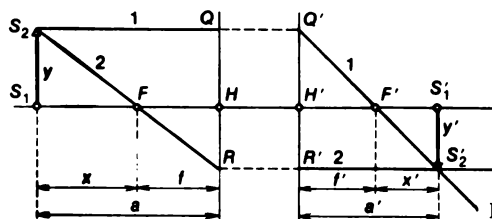


Fig. 226.

Image construction for an optical system.

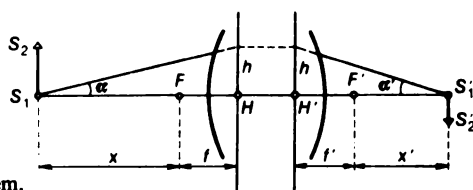


Fig. 227.

Angular magnification of an optical system.

But  $S_1'S_2' = y'$ ,  $H'Q' = HQ = S_1S_2 = y$  and  $F'H' = f'$ . Therefore, denoting by  $x'$  the distance from the rear focus to the image, we get

$$\beta = \frac{y'}{y} = \frac{x'}{f'}. \quad (11.4.1)$$

In the same way, from the similarity of triangles  $S_1S_2F$  and  $HRF$  we obtain

$$\beta = \frac{y'}{y} = \frac{f}{x}, \quad (11.4.2)$$

where  $x$  is the distance from the object to the front focus. (For systems considered in this book (see Sec. 11.2),  $f = f'$ .)

Along with the linear magnification, *angular magnification* also plays an important role in the characteristic of an optical system as in the case of a thin lens (see Sec. 10.10). The angular magnification  $\gamma$  is the ratio of the tangents of angles  $\alpha'$  and  $\alpha$  formed by the rays emerging from the system and incident on it with an optical axis:

$$\gamma = \frac{\tan \alpha'}{\tan \alpha}. \quad (11.4.3)$$

With the help of Fig. 227, we can show (see Ex. 11.45) that

$$\gamma = \frac{1}{\beta} \quad (11.4.4)$$

as in the case of a thin lens.

This means that *the larger the size of the image, the smaller the width of light beams* forming this image (cf. Sec. 10.10). In Sec. 11.11, it will be shown that this circumstance is important for problems dealing with illuminance and brightness of images produced by optical systems.

### 11.5. Drawbacks of Optical Systems

Considering the formation of images of extended objects in optical systems, we always assumed that an image is formed by narrow light beams incident on a system at small angles to its principal axis. Both assumptions are practically never obeyed in optical instruments. For obtaining high illuminance, wide beams have to be used, i.e. large-aperture lenses are required.

The second assumption is also violated in all cases when an instrument must form images of points separated by large distances from the principal

axis (like in photographing). Removing these restrictions, we considerably deteriorate the image: generally, the image becomes blurred and unclear, small details are smudged and become indistinguishable. Besides, the similarity between an object and its image is sometimes lost.

We must also take into account a phenomenon affecting the image formed by an optical system, viz. the dependence of the refractive index of optical glasses on wavelength. This dependence explains why the edges of the image formed in white light are coloured.

In real systems, it is impossible to eliminate completely the drawbacks of optical systems listed above. However, a meticulous analysis of errors of optical systems makes it possible to find ways for reducing their influence. The errors of modern optical instruments are reduced to such an extent that they affect the quality of images insignificantly.

The errors of optical systems are called *aberrations*. Below we shall consider the main types of aberration and methods of their elimination.

### 11.6. Spherical Aberration

The emergence of this error can be traced in simple experiments. Let us take a converging lens *1* (say, planoconvex lens), if possible, with a large diameter and a short focal length. A small but sufficiently bright light source can be obtained by drilling a hole of about 1 mm in diameter in a large screen *2* and fixing in front of the hole a piece of opal glass *3* illuminated by a high-power lamp from a short distance. It would be even better if we focus the light from an arc lamp on this glass. This “luminous point” must be located on the principal axis of the lens (Fig. 228*a*).

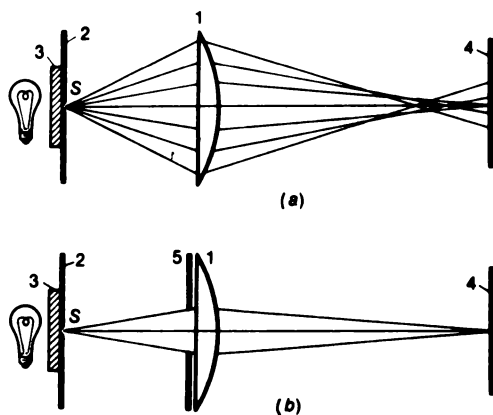


Fig. 228.

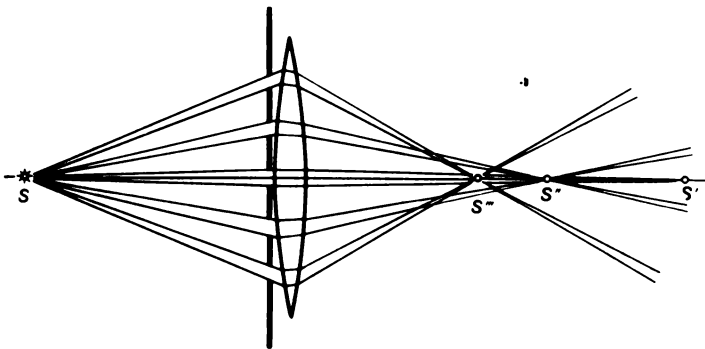
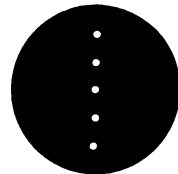
Experimental investigation of spherical aberration: (a) a lens on which a wide beam is incident produces a blurred image, (b) the central zone of the lens produces a sharp image.

Using this lens on which a wide light beam is incident from the glass, it is impossible to obtain a sharp image of the source. Irrespective of the position of the screen 4, a more or less blurred image is formed on it. If, however, the beam incident on the lens is bounded by a piece of cardboard 5 having a small aperture near the central part and placed in front of the lens (Fig. 228*b*), the image will be improved significantly: there exists a position of screen 4 such that the image of the source on it is sufficiently sharp. This observation is in complete agreement with what is known about the image formed by a lens with the help of narrow axial beams (cf. Sec. 10.3).

Let us now replace the piece of cardboard with the central aperture by a piece of cardboard with small holes arranged along the diameter of the lens (Fig. 229). The paths of the rays passing through these holes can be traced by screening with smoke the air behind the lens. We shall see that the rays through the holes at different distances from the centre of the lens intersect the principal axis behind the lens at different points: the further a ray emerges from the axis of the lens, the stronger it is refracted, and the closer to the lens its point of intersection with the principal axis (Fig. 230).

Thus, our experiments show that the rays passing through different zones of a lens arranged at different distances from the axis form images

**Fig. 229.**  
A screen with apertures for studying spherical aberration.



**Fig. 230.**  
Emergence of spherical aberration. The rays emerging from a lens at different heights above the principal axis form images  $S'$ ,  $S''$  and  $S'''$  of point  $S$  at different points.



of the source at different distances from the lens. For a given position of the screen, some of the zones produce sharp images on the screen, while the images formed by other rays are blurred and merge into a bright circle.

As a result, *a large-aperture lens produces the image of a point source in the form of a blurred bright spot rather than in the form of a point.*

Thus, using wide light beams, we cannot obtain a point image even when the point source lies on the principal axis. This error of optical systems is known as *spherical aberration*.

Due to spherical aberration, the focal lengths of the rays passing through the central zone of a simple negative lens are also larger than that for the rays passing through the periphery zone. In other words, a parallel beam passing through the central zone of a diverging lens becomes less diverging than a beam passing through external zones. If we direct the light passing through a converging lens to a diverging lens, we shall increase the focal length. This increase, however, is smaller for central rays than for peripheral ones (Fig. 231).

Thus, a longer focal length of the converging lens, corresponding to central rays, will increase to a smaller extent than a shorter focal length for peripheral rays. Consequently, due to its spherical aberration, the diverging lens levels out the difference in focal lengths for central and peripheral rays, caused by the spherical aberration of the converging lens. This levelling out can be made complete by an appropriate choice of the combination of converging and diverging lenses so that the spherical aberration of the system of the two lenses will practically be eliminated (Fig. 232). Two simple lenses are usually cemented together (Fig. 233).



Fig. 231.

Spherical aberration in a (a) converging lens and (b) diverging lens.

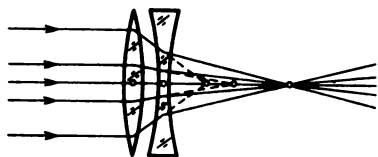


Fig. 232.

Correction of spherical aberration by combining a converging and a diverging lens.

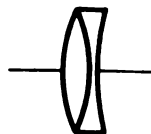


Fig. 233.

Glued astronomical objective corrected to spherical aberration.

It follows hence that spherical aberration can be eliminated by combining two parts of a system, whose spherical aberrations compensate for each other. The other drawbacks of the system are eliminated in the same way.

An astronomical objective may serve as an example of an optical system with eliminated spherical aberration. If a star is on the axis of the objective, its image is practically undistorted by aberration although the diameter of the objective may reach a few tens of centimetres.

### 11.7. Astigmatism

This error of optical systems is manifested during the formation of the image of a point lying at a considerable distance from the principal optical axis of the system. To be more precise, the error arises when we use the light beams which are at considerable angles with the principal optical axis (oblique beams). It is important to note that astigmatism is also observed when narrow beams are employed, and it may be retained in systems free from spherical aberration.

In order to observe astigmatism, we isolate a narrow beam of rays with the help of a cardboard screen having small aperture, and place the source on an auxiliary optical axis forming an angle of  $30\text{--}40^\circ$  with the principal optical axis. It will be seen that the image of a luminous point on screen 4 (see Fig. 228) is rather blurred and has an irregular shape. If we slowly move the screen relative to the lens, we shall find two positions of the screen (*I* and *II* in Fig. 234) in which the image is quite sharp. However, in contrast to the case when the source is on the principal optical axis of the lens, the image in the two positions mentioned above will have the form of a straight segment rather than of a point. The direction of the segment in position *I* is perpendicular to the direction of the segment in position *II*. For all other positions of the screen, the image is blurred (oval or circular).

Therefore, even the best image of a point lying far from the principal axis is not a point but two mutually perpendicular segments located at a distance from each other. This error of optical systems is known as *astigmatism*.

In order to eliminate astigmatism, complex optical systems consisting

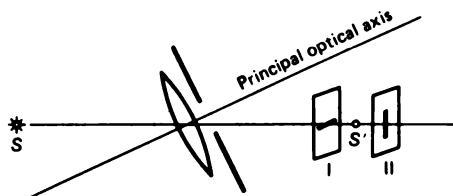


Fig. 234.

Astigmatism of a lens: the images of a point lying on an auxiliary optical axis are two mutually perpendicular segments lying in different planes.

of a few specially selected parts mutually compensating their astigmatism have to be constructed. Systems with eliminated astigmatism are known as anastigmats.<sup>1</sup> Modern photographic objectives with corrected astigmatism produce good images for angles up to  $50\text{--}70^\circ$ .

### 11.8. Chromatic Aberration

Let us first place a red glass (transmitting only red light) and then a blue glass (transmitting blue light) on the path of light rays emerging from lens 1. Using a movable screen 2 (Fig. 235), we shall see that the images formed by the rays of different colours are located at different points:  $S'_r$  (red) is farther from the lens than  $S'_b$  (blue). If we arrange the screen in the position corresponding to the sharp blue image, the blurred spot will be obtained for red rays. For this reason, when white light (containing rays of all colours) is used, the image produced by a lens usually turns out to be coloured (bordered by coloured rings, etc.). This phenomenon is known as *chromatic aberration*.

This error emerges since the refractive index depends on wavelength (dispersion, see Sec. 9.5). As a result, the focal length of a lens, which according to formula (10.2.9) depends on the refractive index, will be different for the rays of different colour. For this reason, the images of point  $S$  for the rays of different colours will be at different distances from the lens.

The distance between points  $S'_b$  and  $S'_r$  depends on the grade of glass of which the lens is made: it is longer for the lens made of glass with a higher dispersion (if lenses being compared have the same focal length for rays of some colour)<sup>2</sup>. This circumstance is used for correcting chromatic aberration of lenses in the following way. A double-convex lens made of glass with a low dispersion is cemented together with an appropriately chosen diverging lens made of glass having a high dispersion (Fig. 236).

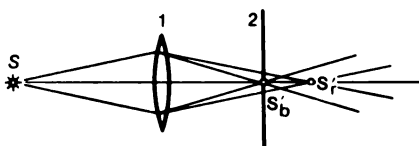


Fig. 235.

Chromatic aberration: the image  $S'_b$  of point  $S$  in blue light does not coincide with the image  $S'_r$  in red light.

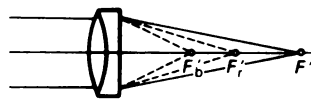


Fig. 236.

<sup>1</sup> The prefix "a" indicates negation. Astigmatism means that the image is not point-like. The prefix "ana" indicates double negation: anastigmatism means nonastigmatism (corrected astigmatism) for which the image is point-like.

<sup>2</sup> I.e. of a glass whose refractive index varies more strongly with a change of the wavelength of incident light.

The additional lens increases the focal lengths of the first lens (see Sec. 11.6) so that the focal length for blue rays, which are refracted stronger, is increased more than the focal length for red rays which are refracted weaker. In the simplest case, the calculation is carried out so that the focus  $F'_r$  for red rays and the focus  $F'_b$  for blue rays coincide at the same point  $F'$ . Superimposed images in different colours will practically yield a white point, i.e. chromatic aberration will be corrected.

Lenses in which chromatic aberration is eliminated by the method described above are known as *achromatic* lenses. The systems in which the foci of three types of rays are combined (*apochromats*) are also used. Such apochromatic systems are employed, for example, in microscopy.

### 11.9. Confinement of Beam Cross Sections in Optical Systems

Studying optical systems, we ignored so far an important circumstance, viz. the limited size of lenses (or mirrors) forming optical systems. This was justified since for constructing images there is no need to know the actual paths of all the rays in a system. For example, to construct the image of a point, it is sufficient to take two rays which generally may not pass through the instrument in actual practice (see Fig. 216).

Since the size of any optical system is limited, most of the rays emerging from a luminous object in all directions pass by the system and cannot participate in image formation. Any obstacle confining the cross section of beams passing through an optical system is called a *diaphragm*. For a simple lens, the diaphragm is just its casing. However, a part of a lens can be covered with a piece of cardboard having an aperture. In this case, the aperture in the cardboard serves as the diaphragm. It should be borne in mind that any part of a lens (provided that the lens is appropriately corrected<sup>3</sup>) forms the same image as the entire lens. Therefore, the presence of a diaphragm does not change the size or shape of the image. It is only the illuminance of this image that changes when a diaphragm is used, since the luminous flux is reduced in this case. If, for example, we cover half a lens with a piece of cardboard, the image remains unchanged, but the illuminance is reduced to half of the previous value since only half of the rays will take part in the image formation.

Thus, the role of the diaphragm for a well-corrected system is mainly reduced to the change of the luminous flux forming the image. The diaphragm also determines the *field of vision* of an instrument, i.e. the maximum part of an object whose image can be formed by the instrument. The role of a diaphragm for obtaining images of extended objects (depth

<sup>3</sup> I.e. the errors indicated above are eliminated.

of focus) will not be considered here. The effect of diaphragm on the resolving power of optical instruments will be considered in Chap. 14.

### 11.10. Lens Aperture

Let us determine the dependence of the illuminance of the image formed by a lens on the characteristics of the lens, viz. on its diameter and focal length.

The illuminance  $E$  of the image is determined by the ratio of the luminous flux  $\Phi$  to the area  $\sigma'$  of the image:  $E = \Phi/\sigma'$ . For a given distance  $a$  from the source to the lens, the luminous flux passing from the source through the lens to the image is proportional to the surface area of the lens, i.e. to  $d^2$ , where  $d$  is the diameter of the lens or of a diaphragm covering the lens. The area of the image is proportional to the squared distance  $a'$  from the image to the lens. If the source is far from the lens, the image is near the focal plane, and the image area is proportional to the squared focal length  $f^2$ . Thus, in the case under consideration the illuminance of the image is proportional to  $(d/f)^2$ .

Indeed, let a surface element  $\sigma$  be located around point  $S$  (Fig. 237) and its image  $\sigma'$  around point  $S'$ . Using the formula for the magnification of a lens, we obtain  $\sigma'/\sigma = a'^2/a^2$ . Further, from the lens equation we have  $\frac{1}{a} + \frac{1}{a'} = \frac{1}{f}$  or  $\frac{a'}{a} = \frac{f}{a-f}$ . If distance  $a$  from the source to the lens is much larger than  $f$ , we can neglect  $f$  in comparison with  $a$  in the denominator on the right-hand side. Then  $a' \approx f$  and  $\sigma'$  is proportional to  $f^2$ .

Thus, *the illuminance of the image formed by a lens is proportional to the square of its diameter and inversely proportional to the square of its focal length*. The quantity  $(d/f)^2$  is known as the *lens aperture*. This quantity characterizes the properties of the lens as regards the illuminance of images formed by it. Sometimes, the quantity  $d/f$  is used instead of  $(d/f)^2$ , known as the *relative aperture* of the lens.

Therefore, the illuminance of the image is reduced when the luminous flux incident on the lens is bounded. This refers to any optical instrument. At the same time, the quality of the image is improved when the bundle of rays is confined.

Thus, it is difficult to combine a good quality of the image and a large lens aperture of the instrument.

In actual practice, a certain compromise has to be made and a loss in the aperture has to be admitted for obtaining an appropriate quality of

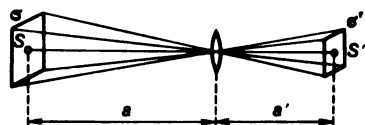


Fig. 237.

To the derivation of the formula for the illuminance of the image formed by a lens.

the image. On the contrary, sometimes we have to bear with the deterioration of the quality of the image for obtaining required illuminance.

In modern optical instruments, large apertures are combined with a sufficiently high quality of images due to the employment of multilens optical systems. In such systems, aberrations introduced by some lenses are compensated by the aberrations due to other lenses. We considered simple examples of correcting optical systems while discussing spherical and chromatic aberrations and astigmatism. It should be noted that the designing of complex optical systems is a complicated problem which requires a high skill and a long time for its solution.

### 11.11. Brightness of Image

It was shown in the previous section that the illuminance of the image of an extended object increases in proportion to the diameter of the lens and in inverse proportion to its focal length. It may seem that the brightness of the image of an extended object can also be increased in this way for obtaining, for example, images which are brighter than the source itself. However, this conclusion is erroneous.

At best, the brightness of the image may be equal to the brightness of source. This is observed in the absence of losses due to partial absorption of light in lenses and partial reflection of light by the surfaces of lenses. In the presence of light losses, *the brightness of an image of an extended object is always lower than brightness of the object itself*. No optical instrument can produce an image of an extended object which would be brighter than the object itself.

It becomes clear why it is impossible to increase the brightness of the image with the help of an optical system if we recall the basic property of any system, mentioned in Sec. 11.4. An optical system operating without losses does not change a luminous flux. But, by reducing the area of the image, it increases in the same proportion the solid angle into which the luminous flux is directed. As the area of the image is reduced, the luminous flux emitted by a unit surface increases, but this flux is directed into a larger solid angle. Thus, *the luminous flux emitted by a unit surface into a unit solid angle, viz. the brightness (see Sec. 8.6), remains unchanged*.

For the simple case of image formation with the help of a lens, we can verify this general conclusion by simple calculations.

Let us arrange a small luminous surface of area  $\sigma$  in front of a lens at a distance  $a$  from it and perpendicularly to the principal axis. Let its image be at a distance  $a'$  from the lens and have the area  $\sigma'$ . Obviously, then  $\sigma/\sigma' = a^2/a'^2$  (Fig. 238), or

$$\sigma/a^2 = \sigma'/a'^2. \quad (11.11.1)$$

Let us calculate the luminous flux passing from the source through the lens. According

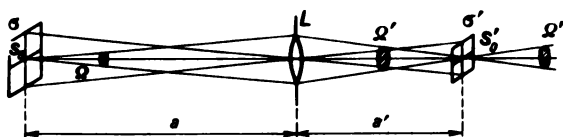


Fig. 238.

Brightness of the image depends on the product of the solid angle and the area of the image and cannot exceed the brightness of the source.

to formula (8.6.2),  $\Phi = L\sigma\Omega$ , where  $L$  is the brightness of the luminous surface element,  $\sigma$  is its area, and  $\Omega$  is the solid angle of the luminous flux directed to the lens. Figure 238 shows that  $\Omega = A/a^2$ , where  $A$  is the area of the lens aperture. Therefore,

$$\Phi = L\sigma A/a^2. \quad (11.11.2)$$

This luminous flux is directed to the image  $\sigma'$ .

The luminous flux emitted by the image is directed into the solid angle  $\Omega'$  which, as is seen from Fig. 238, is equal to  $A/a'^2$ . The flux emitted by the image is equal to  $\Phi' = L'\sigma'\Omega'$ , where  $L'$  is the brightness of the image. Thus,

$$\Phi' = \frac{L'\sigma'A}{a'^2}. \quad (11.11.3)$$

If no light losses occur in the lens, the two luminous fluxes, viz. the one incident on the lens (and directed by it to the image),  $\Phi$ , and the one emitted by the image,  $\Phi'$ , must be equal:

$$\frac{L'\sigma'A}{a'^2} = \frac{L\sigma A}{a^2}.$$

Hence, in view of (11.11.1),

$$L' = L, \quad (11.11.4)$$

i.e. the brightness of the image formed by the lens is equal to the brightness of the object itself. It should be recalled that all conclusions are valid only for extended objects. The brightness of the image of point objects will be considered in the next chapter.

The obtained result allows us to determine the illuminance of the image produced by a lens. According to formula (11.11.3), for the illuminance of the image we can write

$$E' = \frac{\Phi'}{\sigma'} = \frac{L'A}{a'^2}. \quad (11.11.5)$$

If the light losses in the lens can be neglected, then  $L' = L$ , and hence

$$E' = \frac{LA}{a'^2}. \quad (11.11.6)$$

Thus, the illuminance of the image formed with the help of a lens is the same as if the lens were replaced by a source having the same brightness  $L$  and an area equal to the area of the lens surface. The obtained formula (11.11.6) is applicable to more complex systems as well.

The brightness of the image can be increased to the values exceeding the brightness of the source by placing an active medium in the space between the source and the image. This medium will intensify the radiation passing through it. (The methods of obtaining active media

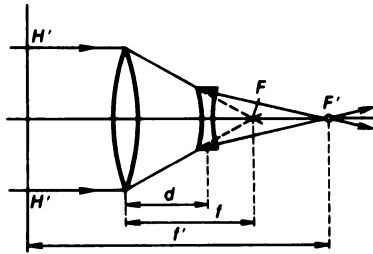
will be considered later.) Systems capable of increasing brightness are referred to as active optical systems. An example of such a system is a laser projection microscope which produces on a screen having an area of a few square metres the images of microscopic objects with an illuminance sufficient for vision in undarkened rooms. In active optical systems, the energy is transferred to the image from an active medium.

? **II.39.** The focal length of an optical system is 30 cm, and the principal planes are separated from each other by a distance of 10 cm. Construct the image of an object located at (a) 20 cm, (b) 50 cm, and (c) 80 cm from the front principal plane. Determine the linear and angular magnifications in each case.

**II.40.** An optical system consists of two lenses arranged in air at 10 cm from each other. The front focus is at 20 cm from the first lens while the rear focus is at 12 cm from the second lens. A three-fold magnified image is at 45 cm from the rear focus. Determine the focal length of the system and the position of the principal planes relative to the lenses forming the system.

**II.41.** Remote objects are photographed with the help of a telescopic objective whose rear principal plane is in front of the first lens (Fig. 239).

Explain the advantages of the telescopic objective over ordinary objective lenses in photographing remote objects.



**Fig. 239.**  
To Exercise II.41.

**II.42.** Determine the relation between the optical power and aperture of a lens.

**II.43.** An object whose illuminance is 40 lx and diffuse reflection coefficient is 0.70 is photographed with the help of an objective with the relative aperture of 1:2.5. Determine the illuminance of the image, assuming that it is nearly in the focal plane of the objective.

**II.44.** Determine the illuminance produced by a floodlight whose mirror has a diameter of 2 m, and the arc has a brightness of  $8 \times 10^8$  cd/m<sup>2</sup> at a distance of 5 km, and the transmission coefficient for air is 0.95.

**II.45.** Prove that for complex optical systems, as well as for thin lenses (see Sec. 10.10), the linear magnification  $\beta$  and the angular magnification  $\gamma$  are related through the formula  $\gamma = 1/\beta$ .

**II.46.** If  $x$  is the distance from the front focus to an object and  $x'$  is the distance from the rear focus to the image, the following relation holds:  $xx' = f^2$  (Newton's formula), where  $f$  is the focal length of the system. Prove the validity of this formula.



## Chapter 12

# Optical Instruments

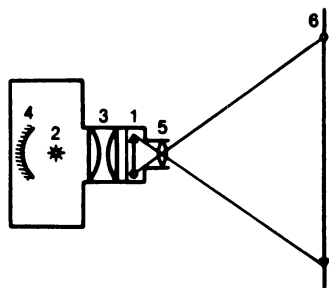
### 12.1. Projection Optical Instruments

The construction of various optical instruments is based on the laws of image formation in optical systems. The basic part of any optical instrument is an optical system. In some optical instruments the image is formed on a screen which has to be arranged in the image plane, while other instruments are intended for operation together with the eye. In the latter case, the eye and the instrument form as if a unified optical system, and the image is formed on the retina of the eye.

We shall consider the operation of optical instruments on the basis of the laws of geometrical optics. However, the concept of light rays turns out to be insufficiently accurate for solving some problems, and we shall have to refer to the wave properties of light which will be studied in the following chapters.

Projectors form a real magnified image of a picture or an object on a screen. Such an image can be observed from a comparatively large distance and hence can simultaneously be seen by a large group of people.

Figure 240 represents the schematic diagram of a projector designed for demonstrating transparent objects like pictures and photographs on glass (slides), films, etc. Such instruments are known as *diascopes* (from the Greek *dia* meaning through). Object 1 is illuminated by a bright light source 2 with the help of a system of lenses 3 called a *condenser*. Sometimes a concave mirror 4 with a light source at the centre is arranged behind the source. The mirror increases the illuminance of the object by reflecting the light incident on the rear wall of the projector back to the system.



**Fig. 240.**

Schematic diagram of a projector for demonstrating transparent objects (transparencies): 1 — object, 2 — light source, 3 — condenser, 4 — concave mirror, 5 — objective, 6 — screen.

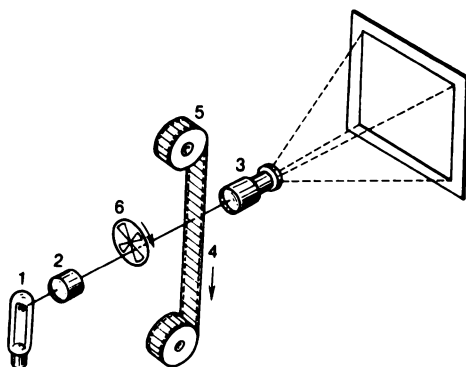


Fig. 241.

Schematic diagram of the simplest cinema projector: 1 — light source, 2 — condenser, 3 — objective, 4 — film, 5 — film-pulling mechanism, 6 — shutter.

An object is arranged near the focal plane of objective 5 which forms an image on screen 6 (see Sec. 10.11). The objective can smoothly be moved for reaching the maximum sharpness of the image.

Projection systems are widely used for demonstrating drawings, figures, etc. while delivering lectures (optical projector). *Cinema projector* is an optical projection system of the same type, the only difference being that demonstrated pictures rapidly follow one another. The film is moved jumpwise frame by frame. During the motion of the film, the light beam is intercepted by a shutter. Figure 241 shows the schematic diagram of the simplest cinema projector.

The image formed by projecting is normally magnified significantly. For example, when a frame of a cinema film, whose size is  $18 \times 24$  mm, is projected on a screen having a size of  $3.6 \times 4.8$  m, the linear magnification is 200, while the area of the image is 40 000 times larger than the area of a frame.

An appropriate choice of a condenser is important for attaining a sufficiently high and uniform illuminance of an object. It may appear that the condenser should concentrate light on the object whose image is to be obtained. This, however, is not true to the fact at all. The attempts to concentrate light on the object usually lead to a strongly diminished image of the light source, and if the latter is not very large, the object will be illuminated nonuniformly. Besides, a part of the luminous flux will pass by the objective of the projector, i.e. will not participate in the image formation on the screen. A proper choice of a condenser allows one to avoid these drawbacks.

Condenser 1 is arranged so that it forms image 6 of a small source 2 on the very objective 3 (Fig. 242). The size of the condenser is chosen

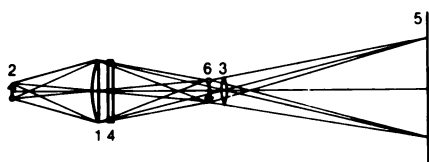


Fig. 242.

Illumination of an object with the help of a condenser: 1 — condenser, 2 — light source, 3 — objective, 4 — slide, 5 — screen, 6 — image.

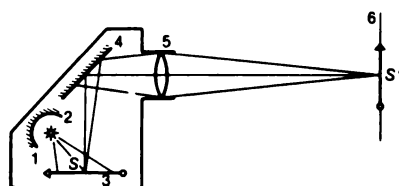


Fig. 243.

Projector for demonstrating opaque objects: 1 — light source, 2 — concave mirror, 3 — object, 4 — plane mirror, 5 — objective, 6 — screen.

in such a way that the entire slide 4 is uniformly illuminated. The rays passing through *any* point of the slide must then pass through image 6 of the light source. Consequently, they will get into the objective and after leaving it will form the image of this point of the slide on the screen.

Thus, *the objective produces the image of the entire slide on the screen. This image correctly represents the distribution of bright and dark regions on the slide.*

In order to demonstrate opaque objects on the screen (like drawings and pictures made on paper), they must be well illuminated from the side with the help of lamps and mirrors and then projected by a fast objective.

The schematic diagram of such an instrument termed *episcopes* (from the Greek *epi* meaning on, upon, to) is shown in Fig. 243. Object 3 is illuminated by source 1 with the help of a concave mirror 2. The rays from each point *S* of the object are turned by a plane mirror 4 and directed to objective 5 which produces the image on screen 6.

Instruments that have a double system for projecting transparent or opaque objects are often used. Such instruments are called *epidiascopes*.

## 12.2. Photographic Camera

The schematic diagram of a photographic camera is shown in Fig. 244. The camera consists of objective 1 and box 2 with opaque walls, which is called a camera. The objective is placed at the front wall of the camera, while a light-sensitive photographic plate 3 is placed at the rear wall. The plate is in a light-tight box (*plate holder*) with a movable lid which is opened

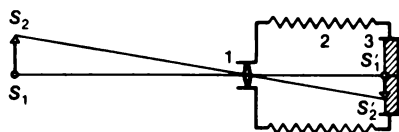


Fig. 244.

Schematic diagram of a photographic camera: 1 — objective, 2 — camera, 3 — photographic plate.

only before shooting. During photographing, the object, as a rule, is at a distance much longer than the focal length of the objective. As a result, an inverted diminished image  $S_1'S_2'$  of object  $S_1S_2$  is formed on the photographic plate (see Sec. 10.11).

In order to obtain a sharp image of an object being photographed, the objective is slightly shifted relative to the rear wall of the camera. For this purpose, the side walls of the first cameras were made in the form of a bellows, the camera as a whole being stretched or compressed. In modern cameras, the adjustment (focussing) is attained by moving the objective in its tube.

The time interval required for illuminating the plate (*exposure time*) depends on the sensitivity of the plate and on the exposure conditions for the object being photographed. Special mechanical shutters are used for shooting with a very short exposure time (of hundredths and thousandths of a second). When the exposure time is long enough, the lid of the camera objective is usually removed for the required time interval.

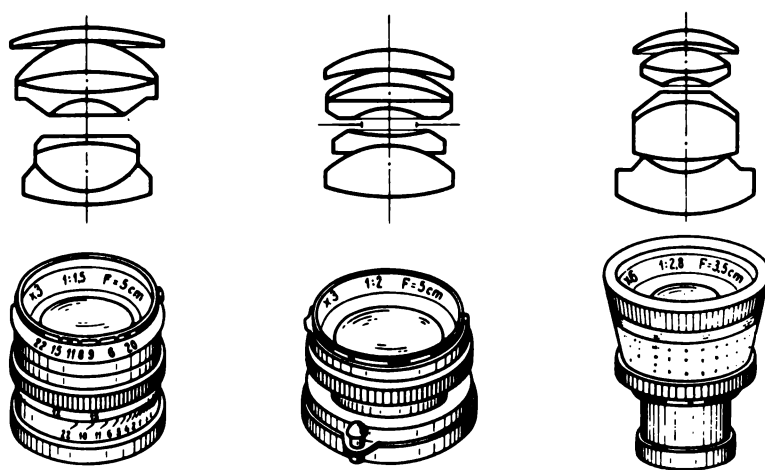
Under the effect of light, the invisible latent image is formed in the light-sensitive layer of a photographic plate. In order to develop this image, the illuminated photographic plate is subjected to a special treatment (see Sec. 21.6).

Depending on duty, diverse constructions of cameras are used. The central part of a camera is its *objective* which mainly determines the quality of the picture and the potentialities of photographing an object under given conditions. In the simplest case, the camera objective is a converging lens. However, it provides a satisfactory quality of the image only at a small aperture ratio and a small angle of vision. Camera objectives that combine a high aperture ratio and a large angle of vision with a high quality of the image normally consist of a few lenses and have a rather complex structure (Fig. 245). Nowadays, the designing of objectives is automated and computerized.

The objective characteristics, i.e. its *focal length*  $f$  (denoted by  $F$  in Fig. 245) and its *aperture ratio*  $d/f$  (see Sec. 11.10) are normally engraved on the lens mount. The aperture ratio is represented in the form of a fraction  $1:a$ , where the quantity  $a = f/d$  is the ratio of the focal length to the lens diameter ( $f$ -number). For example, an objective having a diameter of 20 mm and a focal length of 50 mm has an  $f$ -number of 1:2.5.

The  $f$ -numbers of practically used objective lenses range from 1:7.0 to 1:2.5 for the field of vision of 50-60°. There exist still faster objective lenses (with  $f$ -numbers of 1:1.00-1:0.85).

The luminous flux falling onto a camera is controlled by a diaphragm mounted in the objective. The diameter of the diaphragm, and hence the  $f$ -number, can be varied. The above figures characterize the maximum value of the  $f$ -number of a given objective.



**Fig. 245.**  
Camera objectives (schematic and general view).

It should be noted that the real  $f$ -number of an objective is considerably smaller than the value obtained from a purely geometrical construction. As a matter of fact, the luminous flux incident on a system does not pass completely through it: a fraction of the flux is reflected and another fraction is absorbed in the system. The fraction of absorbed light is usually small, while the role of reflection at the surfaces of lenses is significant. It is well known (see Sec. 9.3) that when light is incident at right angles to a glass-air or air-glass interface, about 4-5% of the incident light is reflected. For an oblique incidence, the fraction of reflected light somewhat increases. Thus, in an objective incorporating three to four lenses, i.e. having six to eight reflecting surfaces, the light losses attain 30-40%.

The reflection of light at the surfaces of lenses not only reduces the aperture ratio but also is responsible for another harmful effect: the reflected light produces a background which masks the difference between bright and dark regions, i.e. reduces the picture *contrast*.

A special technique known as the application of *antireflection coating* has been developed in order to reduce losses for reflection. It consists in the application of a thin transparent film made of appropriate material to the surface of a lens. Owing to interference (see Chap. 13), the fraction of reflected light can be considerably reduced with a proper choice of the film (its thickness and refractive index). The thickness of the layer is usually chosen so that it corresponds to the minimum reflection of green light. Then for shorter and longer waves the reflection is stronger than for green light. When white light is incident on such a surface, the reflected light has a blue-reddish colour. Optical systems in which such surfaces are used are known as "blue" optics. *Such systems have a considerably higher actual aperture ratio and picture contrast than identical optical systems without antireflection coating.*

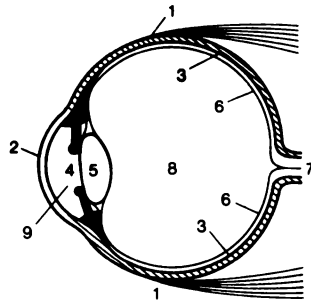
### 12.3. The Human Eye as an Optical System

The human eye has a nearly spherical shape, its diameter being (on the average) 2.5 cm (Fig. 246). The eye is enveloped by three shells.

The outer tough and strong membrane known as *sclera* protects the interior of the eye from mechanical damage. In the front region, the sclera

Fig. 246.

Schematic diagram of the human eye (cross section): 1 — sclera, 2 — cornea, 3 — choroid, 4 — pupil, 5 — lens, 6 — retina, 7 — optic nerve, 8 — vitreous body, 9 — anterior chamber.



is transparent and referred to as *cornea* 2. In the rest of the eye, the sclera is opaque and is known as the *white* of the eye.

The inner surface of the sclera is adjoined by *choroid* 3 consisting of interwound blood vessels feeding the eye. In this front part this second membrane forms the *iris* which is coloured differently for different persons. The aperture in the middle of the iris is called *pupil* 4. The iris can be deformed so that the diameter of the pupil changes. This change occurs reflectively (unconsciously) depending on the luminous flux incident on the eye. The pupil diameter is 2 mm when the illuminance is high and reaches 8 mm when the illuminance is weak.

The inner surface of the choroid contains *retina* 6. It covers the fundus of the eye except its front part. The *optic nerve* 7 emerges from the membrane at the rear part. It connects the eye with the cerebrum. The retina mainly consists of the branched systems of optic nerve fibres and their terminations and forms the light-sensitive surface of the eye.

The space between the cornea and the iris is known as the *front chamber* 9. It is filled with aqueous humour. *Lens* 5 is located in the eye interior, immediately behind the pupil. It is a transparent ductile body in the form of a double-convex lens. The curvature of the surface of the eye lens can be modified by the ciliary muscles of the eye which cover it from all sides. By varying the curvature of the lens surface, the image of the objects lying at various distances can exactly be brought to the surface of the light-sensitive layer of the retina. This process is known as *accommodation*. The entire cavity of the eye behind the lens is filled with a transparent jelly-like liquid forming the *vitreous body* 8.

The structure of the eye as an optical system resembles a photographic camera. The role of the objective is played by the lens together with the refracting medium of the anterior chamber and the vitreous body. The image is formed on the light-sensitive surface of the retina. The image is focussed by accommodation. Finally, the pupil plays the role of the diaphragm.

The ability of the eye to accommodate makes it possible to obtain on

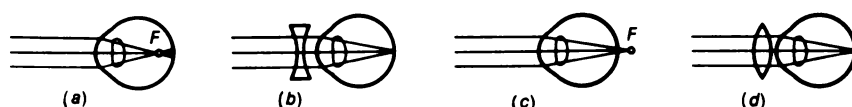


Fig. 247.

Myopia of the eye (a) is corrected with the help of a diverging lens (b), and hyperopia (c), with the help of a converging lens (d).

the retina sharp images of objects lying at different distances. A normal (unstrained) eye, viz. without any effort being made to accommodate, produces a sharp image of remote objects (like stars) on the retina. The eye is focussed to the required distance to the object with the help of a muscular strain increasing the curvature of the lens, and hence reducing its focal length. The shortest distance at which a normal eye can sharply see objects varies depending on age from 10 cm (at an age up to 20 years) to 22 cm (at about 40 years). The ability of the eye to accommodate becomes worse and worse with age: the minimum distance becomes as long as 30 cm (*gerontic hyperopia*).

Not all human beings have normal eyes. Sometimes the rear focus of the unstrained eye is not on the retina proper (like in a normal eye) but on either side of it. If the focus of the relaxed eye lies within the eye in front of the retina (Fig. 247a), the eye is known as *nearsighted*. Such an eye is unable to see remote objects since the strain of the muscles during accommodation moves the focus still further from the retina. For correcting myopia (nearsightedness), the eyes should be supplied with glasses having diverging lenses (Fig. 247b).

For a relaxed *farsighted* eye, the focus is behind the retina (Fig. 247c). The farsighted eye refracts less than a normal eye. For seeing even rather far objects, the farsighted eye has to make no effort. The accommodation capacity of such an eye is insufficient to see close objects. For correcting hyperopia (farsightedness), glasses with converging lenses are being used (Fig. 247d), which bring the focus of the relaxed eye to the retina.

#### 12.4. Optical Instruments Outfitting the Eye

Although the eye is not a thin lens, we can still find in it a point through which the rays pass practically without being refracted, i.e. the point playing the role of the *optical centre* (see Sec. 10.2). The optical centre of the human eye lies in the lens near its rear surface. The distance  $h$  from the optical centre to the retina is known as the *depth of the eye* and normally amounts to 12 mm.

Knowing the position of the optical centre, one can easily construct the image of an object on the retina of the eye. The image is always real, diminished and inverted (Fig. 248a). The angle  $\varphi$  under which object  $S_1S_2$  is seen from the optical centre  $O$  of the eye is called the *angle of view*.

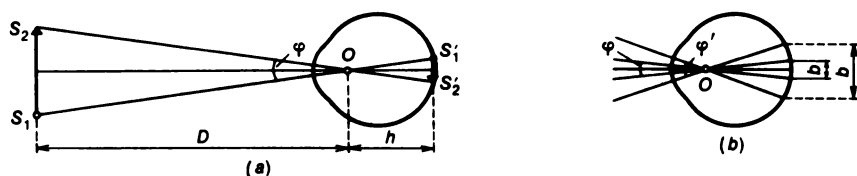


Fig. 248.

(a) Angle of view  $\varphi = S_1'S_2'/h = S_1S_2/D$ . (b) The image of an object formed on the retina increases with the angle of view.  $N = b'/b = \varphi'/\varphi$ .

The retina has a complicated structure and consists of separate light-sensitive elements. Therefore, two points of an object lying so close to each other that their images on the retina get into the same element are perceived by the eye as a single point. The minimum angle of view for which two luminous points (or two black points over a white background) are perceived by the eye as separate is about one angular minute. *The eye poorly recognizes the details of an object seen at an angle less than  $1'$* . This is the angle at which a segment having a length of 1 cm is seen at a distance of 34 m from the eye. At an insufficient illumination (in twilight), the minimum angle of resolution becomes larger and may reach  $1^\circ$ .

By bringing an object closer to the eye, we increase the angle of view, and hence make it possible to resolve finer details. However, objects cannot be brought too close to the eye since it has a limited capacity for accommodation. The most favourable distance for seeing object with a normal eye turns to be 25 cm. At this distance, the eye recognizes details well enough without being tired. This is the *distance of normal vision*. For a nearsighted eye, this distance is somewhat smaller. For this reason, nearsighted persons put the object they want to see closer to the eye than those with the normal vision or farsighted persons so that they see it at a larger angle and can better resolve fine details.

A considerable increase in the angle of view can be attained by using optical instruments. Optical instruments outfitting the eye can be divided into the following two large groups according to their duty.

1. Instruments intended for examining *very small objects* (magnifying glass and microscope). These instruments as if magnify the objects under examination.

2. Instruments intended for observing *far objects* (telescope, binoculars, etc.). These instruments as if "bring the observed objects closer".

Owing to an increase in the angle of view of an object with the help of an optical instrument, the size of the image of the object on the retina increases in comparison with the image formed by the naked eye, and hence the resolving power increases. The ratio of the length  $b'$  of the image on the retina formed by the unaided eye to the length  $b$  of the image formed



by the naked eye (see Fig. 248b) is called the *magnification of the optical instrument*.

Figure 248b shows that the magnification  $N$  is also equal to the ratio of the angle of view  $\varphi'$  of the object viewed through the instrument to the angle of view  $\varphi$  for the naked eye since  $\varphi'$  and  $\varphi$  are small. Thus,

$$N = \frac{b'}{b} = \frac{\varphi'}{\varphi}. \quad (12.4.1)$$

### 12.5. Magnifying Glasses

A magnifying glass, or a *magnifier*, is the simplest instrument for increasing the angle of view. Converging lenses having focal lengths from 10 to 100 mm are used for this purpose. A magnifier is placed in front of the eye as close to it as possible, while the viewed object is at a distance slightly shorter than the focal length of the magnifier. The construction of the image in this case was analyzed in Sec. 10.11. It should be recalled that under these conditions we obtain a virtual erect and magnified image.

Figure 249 illustrates the path of the rays for the case when a small object is viewed through the magnifier. The rays emanating from point  $S$  of object  $l$  are first refracted in the lens of the magnifier and then in the refracting media of the eye, and converge at point  $S''$  on the retina. The rays would have been converged at this point  $S''$  in the absence of the magnifier with the object located at point  $S'$ , i.e. if the eye viewed directly the object of an enlarged size  $l'$  located at the corresponding distance from the eye.

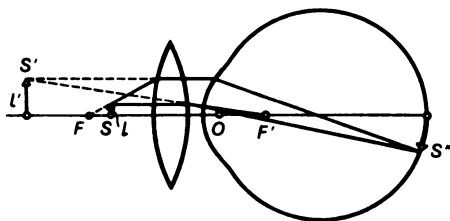
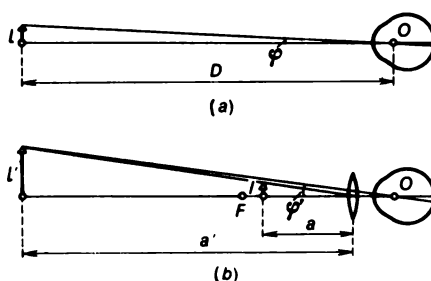


Fig. 249.

Path of rays when a small object is viewed through a magnifier.

The rays shown in Fig. 249 by the dashed lines and intersecting at point  $S'$  form a virtual image of point  $S$  and do not exist in actual practice. An opaque screen placed directly behind the object does not change the situation. However, we "see" object  $l'$  since the eye automatically reconstructs the path of the rays incident on it, and the rays refracted by the magnifier are incident on the eye as if  $l'$  were a real object.

Let us determine the magnification of the magnifying glass. Suppose that an object of length  $l$  (Fig. 250a) is at a distance  $D$  of normal vision

**Fig. 250.**

Examination of a small object by the unaided eye (a) and through a magnifier (b).

from the eye. Then the angle of view is

$$\varphi = \frac{l}{D}.$$

Let us place the same object (Fig. 250b) near focus  $F$  of the magnifier and examine it through the magnifying glass. We shall see the image of length  $l'$  at the angle of view  $\varphi'$  so that

$$\varphi' = \frac{l'}{a'},$$

where  $a'$  is the distance from the magnifier to the image (the distance from the magnifier to the optical centre of the eye is neglected).

According to the formula for the magnification of a lens, we have the following relation:

$$\frac{l'}{l} = \frac{a'}{a},$$

and hence

$$\varphi' = \frac{l}{a}.$$

This gives the following expression for the magnification of the magnifier:

$$N = \frac{\varphi'}{\varphi} = \frac{D}{a}.$$

Since the object is near the focus,  $a \approx f$ . Thus, putting the distance  $D$  of normal vision equal to 250 mm, we obtain the following formula for the magnification of the magnifying glass:

$$N = \frac{250}{f}, \quad (12.5.1)$$

where  $f$  must be expressed in millimetres. For example, for  $f = 50$  mm, the magnifier gives a magnification of 5.

An object can lie just in the focal plane of a magnifier. In this case, a parallel beam of rays will emanate from every point of the object. This beam is converged by the eye to a single point, and a sharp image is formed on the retina of the eye. It should be noted that this case is especially favourable for observations: the normal eye converges the parallel beam to a point, being in the relaxed state. Thus, no efforts are required for accommodation, and the eye does not get tired under such conditions of observation. It is just under these conditions that the magnification of a magnifying glass has the value specified by formula (12.5.1).

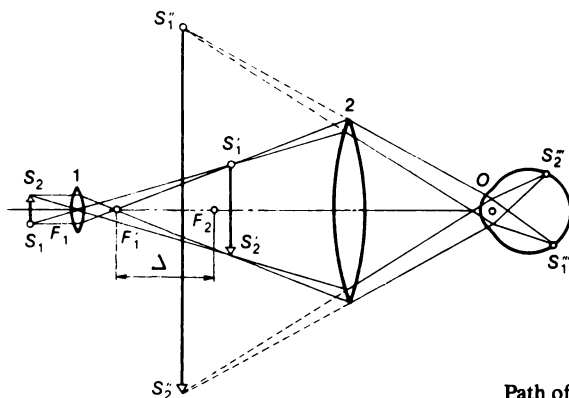
Magnifiers of various types are widely used for doing fine and precision operations, taking measurements, and so on.

It may seem that a magnifier can be used for obtaining very large magnifications when the focal length of the magnifier is made short. For instance, the magnification of a magnifying glass with a focal length of 0.25 mm is 1000. *However, it is practically impossible to employ magnifiers with very short focal lengths, and hence with very small diameters. For this reason, magnifiers with a magnification exceeding 40 are not used.*

## 12.6. Microscopes

Larger magnifications are obtained by using microscopes. The optical system of a *microscope* consists of two parts having more or less complicated structure, i.e. the *objective* (facing an object) and the *eyepiece* (facing the eye). The path of the rays in a microscope is shown in Fig. 251 where the objective and eyepiece are represented just by simple lenses.

Like a magnifier, a microscope makes it possible to view the image of an object at a larger angle than it can be done by the naked eye. A small object  $S_1S_2$  is placed in front of objective 1 of the microscope at a distance slightly exceeding the focal length of the objective. Its real image  $S'_1S'_2$  is near the front focus  $F_2$  of eyepiece 2, i.e. between the eyepiece and the



**Fig. 251.**

Path of rays in a microscope.

front focus. This image is viewed by the eye through the eyepiece as through the magnifier. The image  $S_1''' S_2'''$  formed on the retina is perceived by the eye as that produced by the virtual magnified image  $S_1'' S_2''$  of the object. The distance  $\Delta$  between the rear focus of the objective and the front focus of the eyepiece is known as the *optical length of the microscope tube*. It determines the magnification of the microscope. Image  $S_1' S_2'$  lies in the front focal plane of the eyepiece, which means that image  $S_1'' S_2''$  lies at infinity. In this case, the eye is relaxed.

As for a simple magnifier, the *magnification* of a microscope is equal to the ratio of the length of the image of a segment, obtained on the retina with the help of the microscope, to the length of the image of the segment on the retina of the naked eye.

The operation of a microscope is equivalent to the operation of a simple magnifier with the focal length  $f$  equal to the focal length of the entire microscope. Using formula (12.5.1), we obtain the following expression for the microscope magnification:

$$N = \frac{250}{f}.$$

The focal length of the microscope as a system of two lenses can be made considerably shorter than the focal length of the objective or of the eyepiece taken separately. Accordingly, the magnification of the microscope is considerably higher than those of the objective and of the eyepiece. Calculations show that *the magnification of a microscope is equal to the product of the magnifications of the objective and of the eyepiece*. For this reason, microscopes having a magnification of 1000 and even higher are often used.

The main parts of the optical system of the microscope, viz. objective 1 and eyepiece 2, are placed at the ends of a cylindrical tube mounted on a holder (Fig. 252). Object 3 is placed on stage 4 and illuminated from below with the help of mirror 5 and condenser 6. The mounts of the objective and of the eyepiece are placed in a metal tube 7. Focussing is carried out with the help of the rack screw 8 (rough focussing) or by means of a micrometer screw 9 (fine

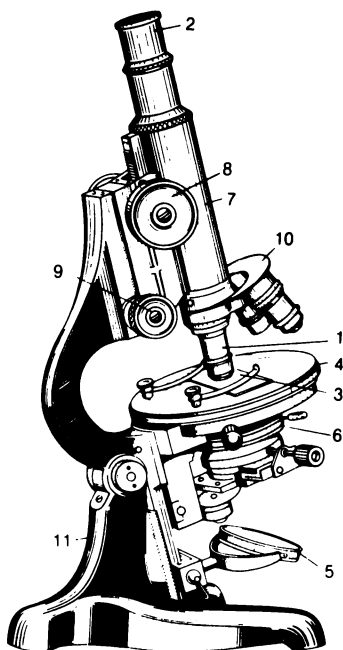


Fig. 252.  
Microscope.

focussing). The magnification of the system can easily be varied since the objectives and eyepieces are accessory lenses. A fast replacement of objectives having different magnifications is carried out by using a revolving nosepiece 10. The tube and the stage are fixed to a heavy holder 11.

The presence of a real intermediate image produced by the objective extends the field of application of microscopes. It allows one to take precise measurements of an object. For this purpose, a scale engraved on a transparent plate is arranged in the focal plane of the eyepiece. The image of the object can be projected on a screen, photographed, and so on (see Ex. 11.53).

### 12.7. Resolving Power of Microscopes

The operation of the microscope was characterized by its magnification.

While discussing the operation of a simple magnifier, it was shown that the magnification provided by an optical system makes it possible to examine a part of an object at a larger angle of view, and hence to discern its finer details. The microscope enables us to distinguish separate details of an object, which are perceived by the naked eye or by means of a magnifier as a point. Thus the fine structure of an object is resolved by the microscope better than by a simple magnifier. However, the *resolving power* of a microscope can be increased by increasing the magnification only to a certain limit. This is due to the fact that our idea about light as rays is too rough, and the wave properties of light must be taken into account. This refers not only to the microscope but also to other optical instruments. The phenomena relating to the wave nature of light will be considered in greater detail later (see Sec. 14.7). Here, it is important to note that the wave nature of light poses a limit on the resolving power of all optical systems<sup>1</sup>, including the microscope. If two points of an object are separated by a distance shorter than a certain limiting value, they cannot be “resolved”: their images will merge irrespective of the magnification of the microscope.

The *maximum resolving power* is attained with as uniform illumination of an object as possible. For this reason, special condensers producing wide beams of rays are employed for illuminating an object. The maximum resolving power is reached when the magnification of a microscope is about 1000.

### 12.8. Telescopes

A *telescope* is an optical instrument intended for viewing rather far objects. Like a microscope, it includes the objective and the eyepiece which are both

---

<sup>1</sup> We mean optical systems for which the concept of resolving power is applicable.

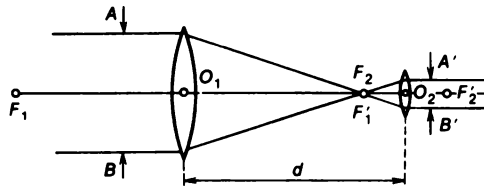


Fig. 253.

Arrangement of the objective and the eyepiece in a telescope: the rear focus  $F'_1$  of the objective coincides with the front focus  $F_2$  of the eyepiece.

more or less complex optical systems (although not as complicated as in the microscope). However, we shall represent them as thin lenses. The objective and the eyepiece of a telescope are arranged so that the rear focus of the objective nearly coincides with the front focus of the eyepiece (Fig. 253). The objective forms in its rear focal plane a real, diminished and inverted image of an object located at infinity. This image is viewed through the eyepiece as through a simple magnifier. If the front focus of the eyepiece coincides with the rear focus of the objective, a parallel beam of rays emanates from the eyepiece when a far object is viewed. This is convenient for observation by a normal relaxed eye (without accommodation, cf. Sec. 12.5). If, however, the eyesight of an observer slightly differs from the normal, the eyepiece is shifted to be arranged according to the eyesight. By moving the eyepiece, a telescope is also focussed for viewing objects located at different but not very large distances from the observer.

The objective of a telescope must always be a converging system, while the eyepiece can be either converging or diverging. The telescope with a *converging* (positive) eyepiece is known as the *Keplerian telescope* (Fig. 254a), while the telescope with a *diverging* (negative) eyepiece is called

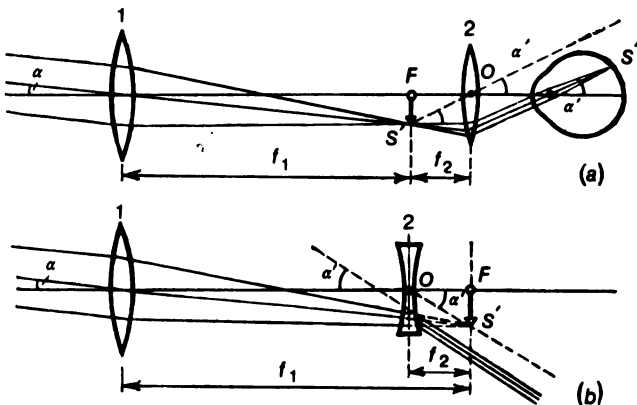


Fig. 254.

Path of rays in a telescope: (a) Keplerian telescope, (b) Galilean telescope.

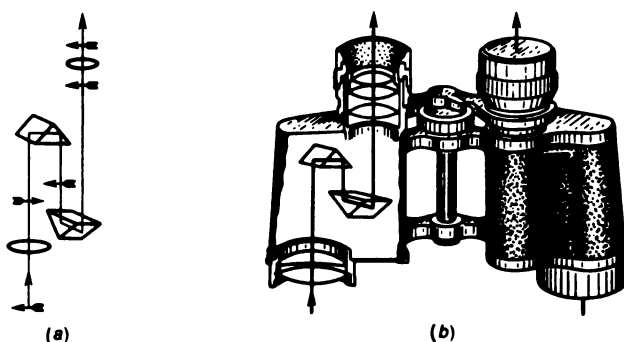


Fig. 255.

Path of rays in a field prism binoculars (a) and its general view (b). Change in the direction of the arrow indicates the "reversal" of the image after passage of the rays through a part of the system.

the *Galilean telescope* (Fig. 254b). Objective 1 of a telescope forms a real inverted image of a far object in its focal plane  $FS'$ . The diverging beam of rays emerging from point  $S'$  is incident on eyepiece 2. As the rays emanate from point  $S'$  lying in the focal plane of the eyepiece, the emerging beam is parallel to the auxiliary optical axis  $S'O$  of the eyepiece at an angle  $\alpha'$  to the principal optical axis. Getting into the eye, the rays converge on its retina and form a real image of the source. (In the case of a Galilean telescope, the eye is not shown in Fig. 254b to avoid the complication of the figure.) Angle  $\alpha$  is made by the rays incident on the objective and the axis.

The Galilean telescope used in conventional opera glasses produces an erect image of an object, while the Keplerian telescope forms an inverted image. For this reason, a Keplerian telescope intended for terrestrial observations is supplied with an *inverting system* (additional lens or system of prisms). As a result, the image becomes erect. An example of such a system is a prism binoculars (Fig. 255). The advantage of the Keplerian telescope is the presence of intermediate real image so that a measuring scale, photographic plate, etc. can be placed in its plane. For this reason, Keplerian telescopes are used in astronomy always when measurements have to be taken.

### 12.9. Magnification of Telescopes

Let  $\varphi$  be the angle at which the rays emerging from the edges of a viewed object get into the eye of an observer in the absence of a telescope (Fig. 256a). (Lens  $L$  forming an image on the screen in its focal plane is shown instead of the optical system of the eye, the screen playing the role of the retina.) By placing a telescope in front of the eye, we increase the angle of view of an object (Fig. 256b). The length of the image on the

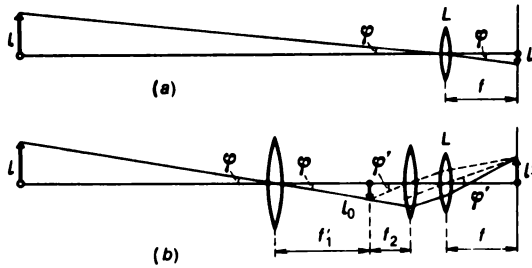


Fig. 256.

Examination of a remote object by the unaided eye (a) and with the help of a telescope (b).

retina (or on the screen in the focal plane of the lens  $L$ ) in the former case is

$$l' = f \tan \varphi \approx f\varphi,$$

while in the latter case it is

$$l'_1 = f \tan \varphi' \approx f\varphi'$$

since angles  $\varphi$  and  $\varphi'$  are small. (Henceforth, we shall write  $\varphi$  and  $\varphi'$  instead of  $\tan \varphi$  and  $\tan \varphi'$ .) The magnification  $N$  of the telescope is

$$N = \frac{l'_1}{l'} = \frac{\varphi'}{\varphi}.$$

Using Fig. 256b, we find

$$l_0 = \varphi f'_1 = \varphi' f_2.$$

Consequently,

$$N = \frac{\varphi'}{\varphi} = \frac{f'_1}{f_2}, \quad (12.9.1)$$

*i.e. the magnification of a telescope is equal to the ratio of the focal lengths of the objective and of the eyepiece.*

Thus, the telescope increases the size of the image of a far object on the retina of the eye, acting so that the object as if is brought closer to the eye. As a result, the eye discerns the details of the object better. Naturally, the resolving power of the telescope is limited by the wave nature of light (see Sec. 12.7).

Various types of telescopes are used as binoculars, in geodesic and military optical instruments, etc.

### 12.10. Telescopes in Astronomy

Telescopes are especially important in astronomy. Galileo Galilei, who was the first to employ a telescope for observing celestial bodies, made a number of important discoveries, although his telescope had only a magnifica-



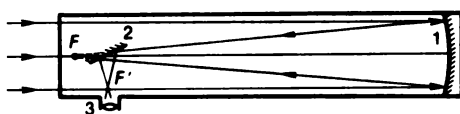


Fig. 257.

Schematic diagram of a reflector telescope.

tion of 30 and provided a rather poor quality of image in our opinion. Modern telescopes are huge in size and have quite a complex structure.

Along with refractor telescopes described above, *mirror-type reflecting telescopes (reflector telescopes)* are of great importance in astronomy.

Figure 257 shows schematically the design of a reflector telescope. Light from a far celestial body is incident on a spherical mirror *1*. Since the light from celestial bodies propagates in the form of nearly parallel beams, the image of a star is formed in the focal plane of the mirror. It is real, inverted and diminished. For the sake of convenience of observation, a small flat mirror *2* is arranged near the focus. It turns the light rays aside. The image formed by the spherical mirror is viewed through eyepiece *3* as through a simple magnifier. The telescope tube protects the mirror from extraneous light.

Let us first consider the image of a comparatively near celestial body (say, a planet) formed by a telescope. The angle of view of a planet viewed by the unaided eye is very small. For example, Mars, whose diameter is 6800 km and which is at  $5.5 \times 10^7$  km from the Earth in the most favourable position, has the angle of view of only  $25''$ . At such a small angle of view, we perceive it as a luminous point. With a telescope, the angle of view of the planet becomes considerably larger, and it is seen as a disc on which some details can be discerned. For example, with a magnification of 75, Mars will be seen through the telescope at an angle of  $31'$ , which is equal to the angle of view of the Sun viewed by the unaided eye.

Stars are at such large distances from us that their details cannot be resolved even when they are observed through the most powerful telescopes. We still see them as luminous points in spite of the fact that the size of some of them is many times larger than that of the Sun. The advantage of the telescope in this case is that the mirror whose area is huge in comparison with that of the pupil receives much more light from the star than the naked eye. Therefore, very weak stars which cannot be seen by the unaided eye can be watched through a telescope. (This question will be clarified in the next section.)

Further, although a telescope forms the images of stars in the form of points, it "moves them apart", which permits astronomers to study the stars which appear as merging when viewed by the naked eye. In other words, the resolving power of a telescope is many times that of the human eye. This question will also be discussed later in the chapter devoted to diffraction.

The capacity of a telescope is determined by the diameter of its aperture. For this reason, the manufacturing of large mirrors and objectives has long become an urgent technological problem. At present, 5-m mirrors are being manufactured. The casting and especially polishing of glasses, as well as the silver plating of such a mirror, are complicated technological problems.

The construction of every new telescope extends the radius of the region of the Universe accessible for observation and widens the possibilities of studying celestial bodies. For instance, a telescope whose diameter is 10 cm makes it possible to observe cracks on the Moon having a width of 40 m and "channels" on Mars 5 km wide. A 5-m telescope resolves "channels" on the Moon whose width is less than 1 m and Martian "channels" just 100 m wide. (Practically, the resolving power of a telescope is slightly lower due to distortions introduced by air flows and imperfections of the optical system.) Thus, the difficulties involved in refining and designing new telescopes are gradually overcome by astronomers and engineers.<sup>2</sup>

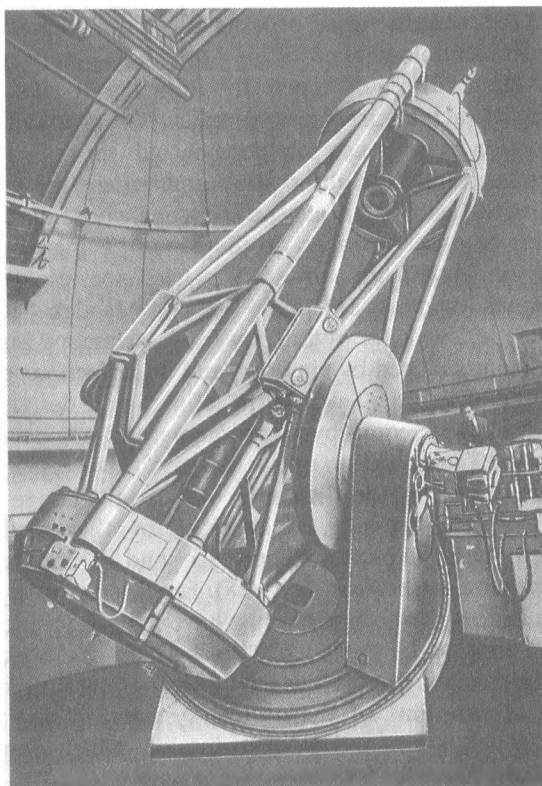
While analyzing the operation of a telescope, it is necessary to consider not only the size of images and  $f$ -number, but also the image quality. Telescopes must provide a high quality of the image. In other words, the optical system of a telescope must be free of spherical and chromatic aberrations, as well as of other faults (see Secs. 11.6-11.8). For this purpose, all refracting and reflecting surfaces of a telescope must have strictly definite shapes correlated with one another, be thoroughly finished, and polished, and so on. For large-scale telescopes, the "correction" of its optical system is a complicated problem. Aberrations of the optical system are eliminated by introducing additional lenses and mirrors into it, which complicates the structure and only partially improves the image.

The other way to improve telescopes consists in giving the mirror the shape of a paraboloid of revolution instead of the spherical shape. A parabolic mirror has a much lower spherical aberration, but its manufacturing presents much more difficulties in comparison with preparing a spherical mirror.

A reflector telescope (Fig. 258) has an advantage over a refractor telescope in that it is free of chromatic aberration. Besides, it is easier to prepare a mirror than an objective since the requirements to the glass homogeneity are less stringent for the mirror because it does not transmit light but is just the base coated by a reflecting layer. For this reason, the largest of the existing telescopes is a reflector whose diameter is 6 m. The diameter of the largest available refractor is only 1 m (with the tube length of 21 m). For the same diameter of 1 m, a reflector should have the tube

---

<sup>2</sup> The operating conditions for telescopes are considerably improved when they are mounted on artificial Earth satellites. In this case, the spectral range of electromagnetic radiation of celestial bodies accessible to observations is widened from the far infrared to the X-ray region.



**Fig. 258.**

A 2.6-m reflector telescope of the Crimean astrophysical observatory of the USSR Academy of Sciences.

length of only 6 m. Therefore, the design of a reflector telescope is simpler than that of a refractor telescope. However, the requirements to the accuracy of manufacturing mirror surfaces are higher than those for objective surface. At the same time, mirrors are more sensitive to bending than lenses. Such bends due to the proper weight of the mirror or as a result of temperature variation considerably deteriorate the image quality. Thus, reflector and refractor telescopes have their advantages and drawbacks.

A successful and ingenious design of a telescope was proposed in 1941 by the Soviet scientist D. D. Maksutov. *In this telescope, the objective is a combination of a positive meniscus* (see Sec. 10.4) *and a mirror*. The positive meniscus must be properly corrected to eliminate the chromatic aberration, but is not free of the spherical aberration. The latter is compensated by a spherical mirror contained in the system and characterized by a spheri-

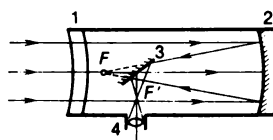


Fig. 259.

Schematic diagram of the Maksutov telescope.

cal aberration of the same magnitude but of the opposite sign. Since the mirror has no chromatic aberration, a system virtually free of both spherical and chromatic aberrations is obtained.

Figure 259 represents the schematic diagram of the Maksutov telescope. In this telescope, a parallel beam of rays, having passed through meniscus 1 and reflected at a concave mirror 2, forms at focus  $F$  an image which is distorted by neither chromatic nor spherical aberration. For convenience of observation, the beam is rotated by a flat mirror 3. The image formed at  $F'$  is viewed through eyepiece 4.

### 12.11. Image Brightness for Extended and Point Sources

In the analysis of operation of optical instruments which operate together with the eye, one should take into account the peculiarities of the eye structure. It was mentioned in Sec. 12.4 that the retina of the eye consists of individual light-sensitive elements each of which responds to light incident on it independently of other elements. When the illuminance of the retina surface increases, only the larger number of its elements take part in perceiving light, the light irritation of each element remaining unchanged. Therefore, *the light perception is determined by the illuminance of the retina*, i.e. by the luminous flux per unit area of its surface. In this respect, the eye resembles a photographic camera where the darkening in a given region of the plate depends on the illuminance of the image in this region and does not depend on the size of the entire image.

Objects having identical brightness and arranged at different distances from the eye will be perceived by it as equally bright. For example, the observed ray of street lights will seem to be uniformly bright, although they are separated by different distances from an observer. To make it clear, it should be recalled that as the distance from a source increases, the luminous flux incident on the eye decreases, but the area of the image formed on the retina decreases in the same proportion. Therefore, the illuminance of the image on the retina, which is equal to the ratio of these quantities, remains unchanged. It is just the illuminance of the retina that determines the light sensation, which thus remains unchanged as the source moves away. (Naturally, it is assumed in this case that the atmosphere is completely transparent, and the removal of the source is not accompanied by an increasing absorption of light.)

The luminous flux received by the eye can be controlled within certain limits (10-15 times) due to the ability of the eye pupil to extend and contract. Under bright illumination, the diameter of the pupil is reduced to 2-3 mm, and the illuminance of the retina drops. On the contrary, when the illumination is poor, the diameter of the pupil increases to the maximum value (7-8 mm), and the illuminance of the retina increases.

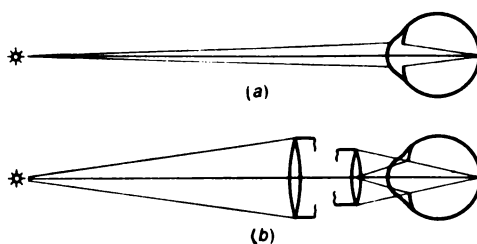
When an optical instrument (like simple magnifier, microscope or telescope) is used for observation, the eye views the image of the object formed by the instrument rather than the object itself. The brightness of the image of an extended object is known to be equal to the brightness of the object itself (see Sec. 11.11) if the light losses in the instrument can be neglected. When the losses due to reflection and refraction of light by refracting surfaces are large, the image brightness is reduced accordingly. Consequently, when extended objects are viewed through an instrument, we never gain in brightness in comparison with the case when the object is viewed by the unaided eye.

An entirely different situation takes place when a point light source like a star is observed. From the point of view of brightness, a point source is a source which is so small that it is seen by the eye at an angle less than  $1'$  under the conditions of normal illumination. When a point source is viewed, its entire image is incident on a single light-sensitive element of the retina. This element receives the entire luminous flux incident on the eye. Therefore, for the point source, the light sensation is determined by the total luminous flux getting to the eye rather than by the illuminance of the retina as in the case of an extended source. Accordingly, the brightness of the image of the point source is determined by the total value of the luminous flux received by the eye.

When a source is viewed by the unaided eye, the luminous flux is proportional to the area of the pupil (Fig. 260a). If, however, the source is viewed, for example, with the help of a telescope, the luminous flux incident on the objective and then on the eye of an observer (Fig. 260b) is proportional to the area of the objective. Therefore, the image brightness increases in proportion to the ratio of the squares of the diameters of the objective and of the eye pupil.<sup>3</sup> For example, using for observations a telescope with an objective diameter of 60 cm and putting the diameter of the eye pupil equal to 6 mm, we have a 10 000-fold increase in the brightness of stars. It is remarkable that the brightness of the sky background as of an extended object does not increase during observations through a telescope. This means that the contrast of the image is improved considerably. Owing to this, stars can be seen through the telescope even in the daytime.

---

<sup>3</sup> It is assumed here that the objective and the eyepiece of a telescope are chosen correctly so that the entire luminous flux incident on the objective gets into the pupil of an observer.

**Fig. 260.**

When a telescope is used, the luminous flux incident on the eye is considerably augmented.

Thus, in contrast to the case when extended bodies are observed, optical instruments make it possible to see weak point sources much better. It is for these reasons that powerful telescopes having a diameter of several metres allow us to see stars whose brightness is about  $10^{-4}$  of that perceived by the unaided eye.

### 12.12. Lomonosov's Telescope

Although a telescope does not increase the image brightness of an extended body observed through it, the ability to discern details of weakly illuminated objects is considerably improved. This fact was established for the first time by the great Russian scientist M.V. Lomonosov (1711-1765), the founder of science and instrument making in Russia. He constructed a "night-vision" telescope intended for the observation of objects in twilight or at night, when the illumination is weak. In these conditions, the properties of a perceiving eye considerably differ in comparison with the daytime conditions. In particular, as was mentioned earlier (see Sec. 12.4), the minimum angle of view for which the eye can distinguish between two points of an object becomes much larger. A telescope increases the angle of view of objects and improves the ability to discern objects under the conditions of night or twilight illumination.

The optical system of a "night-vision telescope" is made of the minimum possible number of parts in order to reduce the light losses due to reflection and absorption. For the same purpose, the surfaces of the optical elements are coated (see Sec. 12.2). The telescope must be designed for the operation with the maximum diameter of the eye pupil (about 8 mm). Besides, the "night-vision telescope" must have the maximum possible  $f$ -number and a large magnification.

### 12.13. Binocular Vision and Sensation of Depth. Stereoscopes

Visual perception of external space is a complex action in which it is essential that normally two eyes participate in vision. The same object is projected on the retinas of both eyes, the two images being slightly different since the object is located not exactly in the same way relative to the eyes: one eye sees slightly better the right side, while the other eye sees better the left side of the object. This difference is negligibly small when a flat object (say, a picture) is viewed, but becomes noticeable in observing three-dimensional objects. Light irritations of each eye are combined in our consciousness into a single visual image in which the peculiarities due to the spatial nature of the observed object are reflected.

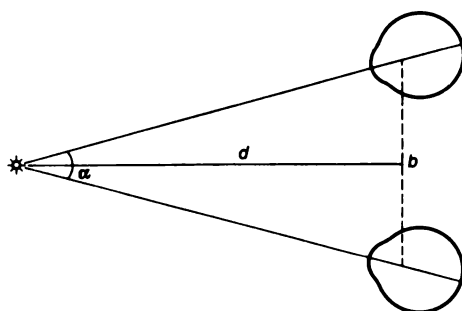


Fig. 261.

Examination of an object by two eyes makes it possible to estimate the distance to the object. Angle  $\alpha$  in the figure is much larger than in actual practice when extended objects are being examined.

In the attempt to examine an object, we rotate the eyes so that their optical axes intersect at the object (Fig. 261). Owing to the high mobility of eyes, we rapidly fix successively different points of the object. In this process, it is possible to estimate the distances to the objects under examination and compare these distances. Such an estimate gives an idea about the *depth of the space* (perspective) and about the spatial distribution of details of the object under investigation or, in other words, makes a *stereoscopic vision* possible.

While viewing with a single eye, we also estimate the relative arrangement of objects by using indirect indications: comparing the dimensions of an object with those of other objects known to us from experience, analyzing the change in colour and location of light and shadows, the superposition of contours of objects, and so on.

The observation of the relative displacement of objects with movements of an observer eye also turns out to be helpful. Besides, distances can be estimated from the sensation of muscular tension required for the accommodation of the eye to a given object. For binocular vision, we also use the sensation of the muscular effort required for the arrangement of the optical axes of the eyes to a fixed point. The two latter processes occur simultaneously and subconsciously, being closely related to each other.

The spatial depth in binocular vision is perceived much better than with a single eye. To prove this, it is sufficient to close one eye and try to thread a needle. The sensation of spatial depth and the ability to estimate the relative displacement of objects from the depth is different for different persons and depends, for one, on the training.

The angle  $\alpha$  of divergence of rays emitted by a far object to the eyes is proportional to the distance  $b$  between the eyes (called the *base*) and

is inversely proportional to the distance  $d$  to the object (see Fig. 261):

$$\alpha = \frac{b}{d}.$$

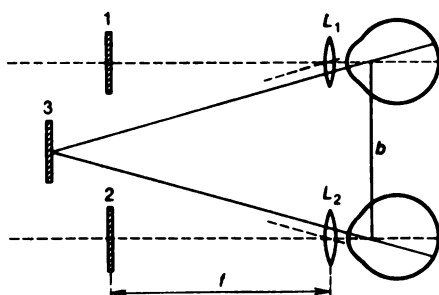
For large distances from the object, the angle  $\alpha$  is very small, and the optical axes of the two eyes are almost parallel. As a result, the sensation of the spatial depth is lost. This angle can be considerably increased by using optical instruments intended for observations at the expense of increasing the base between the objectives of an instrument. Owing to this effect, the sensation of the depth increases many times.

In military optical instruments intended for observations (binoculars and stereoscopic telescopes), the distance between the centres of the objectives is always considerably longer than the separation of the eyes, and hence far objects always seem to be much more contrast in comparison with the case when they are viewed by the naked eye. On the contrary, opera glasses are intended for observing the stage whose real depth is small and the sensation of depth is produced artificially, with the help of scenery. For this reason, the distance between the objectives in opera glasses is made shorter than the distance between the eyes. As a result, the small depth of the stage becomes less noticeable. Naturally, such an arrangement of objective lenses is possible only in prism binoculars where the presence of the prisms enables us to make the distance between the objectives other than the distance between the eyepieces (eyes). Figure 255 shows field prism binoculars with an increased base.

The role of two eyes can be played by two photographic cameras whose optical axes are parallel and displaced relative to each other by a distance  $b$  and which are usually combined into a single camera with two objectives. Naturally, instead of two cameras we can use one camera and photograph an object with it consecutively from two different positions. If we arrange the obtained pictures so that the right eye sees only the picture taken by the right camera and the left eye, that taken by the left camera, and appropriately direct the axes of the eyes, the images of the two pictures will be combined, and the observer will see a relief space. Indeed, in these conditions the images of these pictures on the retina are similar in their geometrical properties and arrangement to the images of the real object viewed directly by the two eyes.

An instrument called *stereoscope* simplified the examination of pictures obtained by the above method. The schematic diagram of a stereoscope is shown in Fig. 262. Stereoscopic photographs 1 and 2 are viewed through lenses  $L_1$  and  $L_2$  each of which is arranged in front of an eye. The photographs are arranged in the focal planes of the lenses, and hence their images lie at infinity. Both eyes are accommodated to infinity. The images of the photographs are perceived as a single relief object lying in plane 3.





**Fig. 262.**  
Schematic diagram of a stereoscope.

Stereoscopes are widely used now for analyzing photographs of country. Photographing an area from two points, one obtains two photographs. Examining them through a stereoscope, one can easily see the relief of the country.

The high sharpness of stereoscopic vision makes it possible to employ a stereoscope for revealing forged papers, money, etc. Examining through a stereoscope a bank note and its forged copy, we shall see a relief instead of a single flat image since all nonidentical details of the objects being compared produce the impression of relief details projecting over a flat background.

- II.47.** The objective of a projector has a focal length of 20 cm. What must be the distance between a  $9 \times 12$  cm slide and the objective for its image to exactly fit into the screen whose size is  $3 \times 4$  m?
- II.48.** The airborne prospect from the height of 3000 m must yield photographs of the country to scale 1:5000. What must be the focal length of the objective?
- II.49.** Determine the reflection losses in the periscope of a submarine having 40 reflecting surfaces, assuming that 5% of light is lost at each surface. Determine the losses in the same periscope with coated lenses, assuming that 1% of incident light is lost at each surface after coating.
- II.50.** A spectacle lens having a focal power of +8 D is used as a magnifying glass. Determine its magnification.
- II.51.** Calculate the maximum diameter of a plano-convex lens having a spherical surface made of glass with a refractive index of 1.63 and used as a simple magnifier producing a 200-fold magnification. (The lens under consideration is not thin. However, this circumstance can be neglected in order to simplify calculations.)
- II.52.** Derive the formula for the magnification of a simple magnifier for the case when an observer arranges it at the distance of normal vision.
- II.53.** How can the image produced by a microscope be projected on a screen?
- II.54.** The magnification of a microscope with a  $7\times$  eyepiece is 140. What will be the magnification of the microscope if its eyepiece is replaced by a lens with a focal length of 10 mm?
- II.55.** Show that the optical system shown in Fig. 255 (a prism binocular) indeed forms an erect image.
- II.56.** Determine the magnification of a reflecting telescope whose mirror has a radius of curvature of 2 m and the focal length of the eyepiece is 20 mm.
- II.57.** Determine the magnification of a telescope having an objective with a focal length of 1600 mm and a  $10\times$  eyepiece.

**II.58.** Given two positive lenses with focal lengths of 3 cm and 15 cm. How should they be arranged to obtain a telescope? Which of the lenses will play the role of objective? What will be the magnification of such a telescope?

**II.59.** What must be the lenses for designing a  $12\times$  Galilean telescope having a length of 22 cm?

**II.60.** The objective of a projector has a focal length of 25 cm and a diameter of 20 mm. A  $6 \times 9$  cm slide is at 26 cm from the objective. Determine the size of the screen and its separation from the objective. Design the condenser (i.e. determine its diameter and focal length) if the light source is an arc with a crater diameter of 9 mm. Determine the distance from the source to the condenser and from the condenser to the projector objective.

**II.61.** The objective of a camera has a focal length of 50 mm. What must be the exposure time for photographing a motor car moving at a distance of 2 km at a velocity of 72 km/h for the displacement of the image during shooting to be less than 0.005 mm?

## Part Three

# Physical Optics

### Chapter 13

## Interference of Light

#### 13.1. Geometrical and Physical Optics

It follows from what has been said in Part Two of the book that a wide circle of problems in practical optics can be solved without referring to wave concepts of light. For this purpose, we introduced the concept of light ray as the line indicating the direction of propagation of luminous energy. Besides, we established geometrical rules concerning the variation of the direction of light rays as a result of reflection or refraction of light. Using these rules, we considered in Chapters 10-12 a large number of important problems in practical optics. The problems that can be solved geometrically constitute geometrical, or ray, optics.

However, while considering these problems, we come across important questions concerning the resolving power of optical instruments, that cannot be answered in the framework of geometrical optics. Moreover, there exists a vast class of optical problems mainly related to the interaction between light and matter which require a deeper knowledge about the nature of light for their solution. These problems are considered in the so-called *physical optics*. This chapter will be devoted to the fundamentals of this branch of physics.

#### 13.2. Experimental Realization of Interference of Light

Colours of thin films described in Sec. 7.2 is the most abundant and easily observed case of *interference*. However, the conditions of emergence of interference pattern in this case considerably differ from the conditions under which interference of waves on the surface of water is observed (see Sec. 5.2). In experiments with waves on the water surface we had two sources of waves (two points), while for the interference in thin films we have only one source of light. Then where do two interfering waves come from? Is it possible to realize interference of light if we have light waves emitted by two different sources like two incandescent lamps or two regions of a red-hot body? The answer to the latter question is provided by everyday experience. It is well known that *interference is not observed when the same region is illuminated by light from different sources*. If two electric bulbs are switched on in a room, the light from one source enhances the illumina-

tion produced by the other source in the entire illuminated region, and the addition of the second source does not lead to the formation of maxima and minima of illuminance.

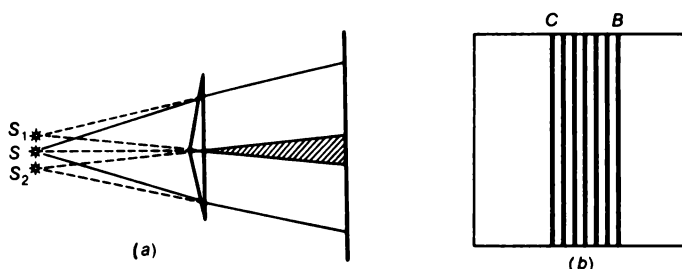
The reason behind this fact is that, as was mentioned in Sec. 5.2, the *coherence* or matching of two systems of waves must be ensured for obtaining a stable interference pattern. Sources must emit coherent waves, i.e. the waves having the same period and a constant phase shift over the time interval sufficiently long for observation. Our ways of observation (by using the eye, photographic plate, etc.) all require comparatively long intervals of time measured in thousandth and even smaller fractions of a second. In independent sources, light is emitted by different atoms for which radiation conditions change rapidly and chaotically. The latest data show that such changes occur at best in about  $10^{-8}$  s and usually much more frequently. Thus, the interference pattern obtained from independent sources remains unchanged for a very short time and is then replaced by another pattern characterized by a different arrangement of maxima and minima. Since, as was mentioned above, the observation time is of the order of  $10^{-3}$  s (or more), interference patterns replace one another over this interval millions of times. We observe the result of superposition of these patterns. Obviously, such a superposition blurs the pattern without leaving any traces of interference fringes. Thus, it is clear why no interference pattern is observed when we have two independent incoherent sources of light. However, interference can be observed if we have two different *laser* sources of light.

In order to observe interference, we must resort to an artificial technique. It consists in that we make interfere the parts of the same wave emitted by a single source and reaching the point of observation via different routes. As a result, a certain path difference appears. The coherence is ensured by the fact that the two interfering waves are emitted by the same source. In experiments with thin films, the wave emitted by a source is split into two as a result of reflection from the front and rear surfaces of the film.

The same result can be attained by using other techniques, say, with the help of the so-called *biprism*<sup>1</sup> (Fig. 263a), where the wave bifurcation is attained due to refraction. The situation is such as if two coherent sources were at points  $S_1$  and  $S_2$ . Actually, we have a single real source  $S$ . This source is a narrow illuminated slit parallel to the edge of the biprism. The wave propagating from the source is split into two parts as a result of refraction in two halves of the biprism and arrives at the points of a screen via two different routes, i.e. with a certain path difference. A system of alternating bright and dark fringes parallel to the edge of the biprism

---

<sup>1</sup> The prefix "bi" is from the Latin *bis* meaning double; biprism means a double prism.

**Fig. 263.**

Observation of interference of light with the help of Fresnel's biprism: (a) schematic diagram of the experiment (top view) and (b) the interference pattern.

will be observed on the screen (Fig. 263*b*). The fringes are located in the region of the screen where the light beams passing through the two halves of the biprism overlap (hatched region in Fig. 263*a*).

The path difference between the two interfering rays is limited due to the following considerations. During each radiation act, an atom emits a system of waves (wave train) propagating in space and time and retaining its sinusoidal nature (see Sec. 1.5). The duration of a train, however, is limited due to the attenuation of electron's oscillations in the atom and due to collisions of this atom with other atoms. The length of a train, or, as it is called, the coherence length of the train, attains 30 cm in the most favourable conditions, while its duration does not exceed  $10^{-8}$ - $10^{-9}$ . The necessary condition for the interference is that the propagation difference (the difference in optical paths, i.e. the products of the refractive indices and geometrical path lengths) for the two rays do not exceed the coherence length of the wave train generating these rays. Figure 264 illustrates this condition.

An ideal source of light is a quantum oscillator (laser) which is coherent by nature as a source of induced radiation (see Sec. 22.9). The coherence length of a wave train emitted by a laser amounts to thousand kilometers, the duration of the train reaching hundredths of a second. Thanks to quantum oscillator, it has become possible to develop a new branch of optics, viz. coherent optics, which has demonstrated brilliant achievements in the theory and technology and whose prospects are virtually unlimited.

If the source of light in the experiment with a biprism (Fresnel's experiment) emits white light, a coloured interference pattern will be observed on the screen as in the case when interference is observed in thin films. If, however, the source emits a monochromatic light (of the colour) like the light from an arc discharge in a gas, passed through an appropriate light filter, the interference pattern is formed by alternating light and dark fringes. The arrangement of these fringes depends on the colour, so that

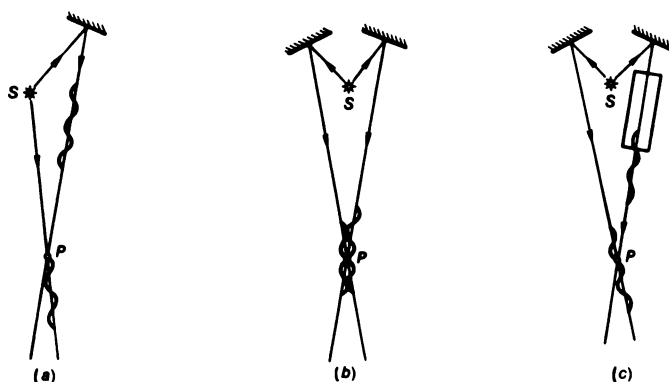


Fig. 264.

To the interference of two trains of light waves: (a) the propagation difference of the two trains is larger than the coherence length, and there is no interference, (b) the propagation difference is zero, and interference is observed, (c) a glass plate is placed on the way of one of the trains ( $n > 1$ ), the propagation difference of both trains is larger than the coherence length, and there is no interference.

regions corresponding to a minimum in one colour may turn out to be regions of maximum in another colour. This means that the distance from sources  $S_1$  and  $S_2$  to the region on the screen under examination is expressed by an even number of half-waves for one colour and by an odd number of half-waves for the other colour. In other words, wavelengths corresponding to different colours are different. Thus, *light of different colours is characterized by different wavelengths.*

Since the position of interference fringes is determined by the wavelength, Fresnel's experiment can be used for determining the length of a light wave if appropriate measurements are made. Such measurements showed that, for example, a flame coloured by sodium vapour (yellow colour) emits light characterized by two wavelengths, viz. 589.0 and 589.6 nm. Measurements show that *the wavelength decreases upon a transition from red to violet in the order of arrangement of colours in a rainbow.*

It is known that the estimate given for a colour by eye is rather indefinite so that the words "red" or "yellow" colours pertain to different shades. Therefore, the wavelength indicated for each colour is of tentative nature. Violet corresponds to wavelength from 400 to 450 nm, indigo from 450 to 480 nm, blue from 480 to 500 nm, green from 500 to 560 nm, yellow from 560 to 590 nm, orange from 590 to 620 nm, and red from 620 to 760 nm. Thus, indicating a colour we characterize light only approximately. On the other hand, the wavelength is an exact quantitative characteristic of a colour, which is used in all scientific measurements.

### 13.3. Explanation of Thin Film Colours

On the basis of what was said in the previous section, we can form a clear idea about the origin of colours in thin films. When a transparent film is illuminated, a part of light wave is reflected from the front surface, and the other part is reflected from the rear surface. As a result, two waves with a certain propagation difference are superimposed. It can easily be seen (Fig. 265) that this propagation difference depends on the film thickness which determines the path length of the wave within the film. In the

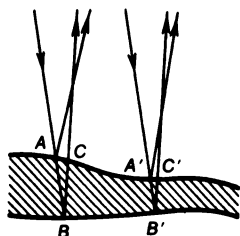
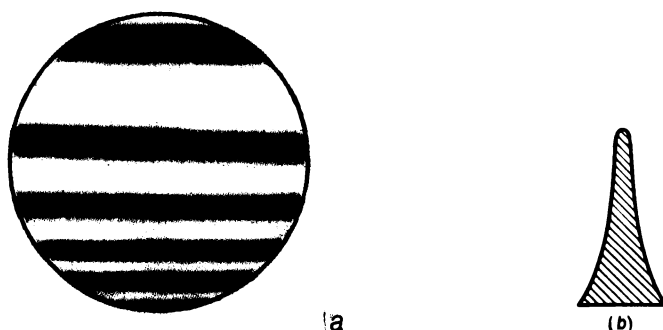


Fig. 265.

Propagation difference ( $ABC, A'B'C'$ ) for two portions of a light wave reflected from the front and the rear surface of a thin film depends on the film thickness in the region of reflection.

regions where this propagation difference is equal to an even number of half-waves, the two parts of the wave enhance each other (maximum), while in regions where the propagation difference is expressed through an odd number of half-waves, they suppress each other (minimum). Since the film may have different thickness in different regions, the regions of maxima and minima form on its surface a pattern of light and dark fringes if the experiment is carried out in a monochromatic light and coloured fringes if white light is used for the experiment. In order to observe this interference pattern, one should examine the surface of the film, i.e. the eye should be accommodated to its surface. This means that the interference pattern is localized near the surface of the film. In some cases, this can be revealed by moving a miniature light receiver (photocell or thermocouple) connected to a galvanometer near the surface of the film. Maximum and minimum readings of the galvanometer alternating as the photoreceiver is moved, confirm the nonuniform distribution of illuminance in the interference light field near the film. The pattern of interference fringes in such experiments indicates the distribution of regions having the same thickness in the field and gives an idea about the shape of the film. For example, Fig. 266 shows that the film has the form of a vertical wedge. Such a film can be prepared by immersing a wire ring into a soap solution and arranging it in the vertical position. Under the action of gravity, the solution flows down, and the film acquires the shape of a wedge flat at the top and gradually expanding to the bottom (Fig. 266b). Examining such a wedge in the light from the Sun or from a projector, we shall see a series of horizontal coloured fringes parallel to the edge of the wedge. Fringes are repeated in a certain sequence of colours. In a monochromatic light (red light filter), we shall obtain alter-

**Fig. 266.**

Interference fringes (a) in a wedge-shaped film (b): the width of the fringes decreases in the downward direction with increasing thickness of the film. The film thickness is exaggerated. The film thickness does not exceed a few micrometres even at the bottom.

nating bright (red) and dark fringes of the same shape (see Fig. IV on the end leaf). In films with randomly distributed thickness (like an oil film on the water surface), the arrangement of maxima and minima is intricate. The role of the angle at which a film is viewed is also obvious. Depending on the direction to the observer, and the angle of incidence of light to the film, the length of the path of light within the film will be larger or smaller, and hence the propagation difference for the parts of the wave that are reflected from the front and rear surfaces will be different.

### 13.4. Newton's Rings

As was mentioned above, the intricate form of interference fringes in thin films is explained by random nonuniformities in the film thickness. In a wedge-shaped film, the regions of equal thickness are arranged along the edge of the wedge. Accordingly, dark and light (coloured) interference fringes are also arranged similarly.

An important modification of the experiments with a wedge-shaped film was the one carried out in 1675 by the English physicist and mathematician Isaac Newton (1643-1727). He observed the colours of a thin air film contained between a flat glass plate and a convex surface of the objective lens of a refracting telescope. The radius of curvature of the objective surface in Newton's experiment was about 10 m, and hence the thickness of the air film between tightly pressed glasses gradually and regularly increased from the point of contact between glasses (where it was zero) to the edges of the objective.

An analysis of interference pattern shows that the dark spot of contact between the flat plate and the lens is surrounded by an annular light fringe

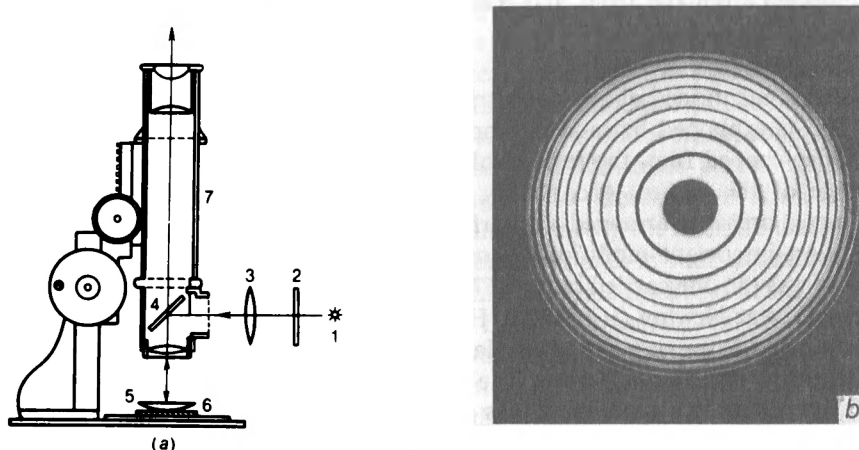


which gradually becomes dark, is alternating with a light ring, and so on. The thickness of the air film increases with the diameter of rings nonuniformly, and the air wedge becomes steeper and steeper. Accordingly, the width of annular fringes, i.e. the distance between two adjacent minima, becomes smaller and smaller. Such a pattern is observed in a monochromatic light. In white light, a system of coloured rings gradually changing one another is observed. With increasing distance from the central dark spot, the coloured fringes become narrower and whiter due to overlapping of colours until any trace of interference pattern disappears.

On the basis of what was said above, it is not difficult to explain why the interference pattern has the form of a system of concentric rings. The regions of equal thickness in the air film, which correspond to the regions of the same propagation difference of light waves, have the form of rings. The regions of the same illuminance in the interference pattern coincide with these rings.

Experimental arrangement for observing and measuring Newton's rings is shown in Fig. 267.

A flat glass plate in contact with a lens having a small curvature is placed on the stage of a microscope with a small magnification. The pattern is viewed through the microscope in the direction perpendicular to the plane of the glass plate. To make light fall at right angles to the surface of the plate, it is passed from the source to be reflected at a glass plate arranged at  $45^\circ$  to the microscope axis. Thus, the interference pattern is observed



**Fig. 267.**

Observation of Newton's interference rings: (a) schematic diagram of the experiment, (b) interference rings. 1 — light source (a lamp with filter 2, or a sodium burner), 3 — auxiliary condenser, 4 — glass plate reflecting light, 5 — long-focus lens and 6 — plane plate forming together an air wedge, 7 — microscope for observing the rings and measuring their diameter.

through this glass plate. It practically does not deteriorate the observation of the rings since the light flux passing through it is sufficient for observation. To improve the illuminance, a condenser can be used. The source of light is a burner whose flame is coloured by sodium vapour (monochromatic light), or an incandescent lamp which can be covered by a light filter.

### 13.5. Calculation of Wavelength of Light with the Help of Newton's Rings

In order to make use of interference phenomena, in particular, Newton's rings, for measuring wavelength, we must consider in greater detail the conditions for the formation of illuminance maxima and minima.

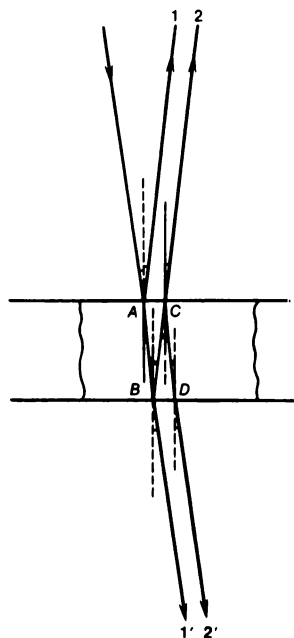
When light is incident on a film or thin plate, it partially passes through it, and a part of it is reflected. Let us assume that a monochromatic light with a wavelength  $\lambda$  is incident on a plate at right angles to its surface. We shall consider a small region of the plate, assuming that it is plane-parallel. Figure 268 shows the paths of rays in the plate. For the sake of convenience, the rays are drawn so that they are not strictly perpendicular to it. In the reflected light, we have ray 1 reflected from the upper surface of the plate and ray 2 reflected from the lower surface. In the transmitted light, ray 1' passes through the plate and ray 2' is reflected once from the lower and once from the upper part of the plate.<sup>2</sup>

Let us first consider transmitted rays. Rays 1' and 2' have a propagation difference since the former has passed through the film once while the latter three times. For the normal incidence, the propagation difference formed is  $AB + BC + CD - AB = BC + CD = 2h$ , where  $h$  is the thickness of the plate. If this difference is equal to an integral number of wavelengths, i.e. an even number of half-waves, the rays enhance each other. If the difference is equal to an odd number of half-waves, the rays are mutually suppressed. Thus, the maxima and minima are formed in the regions of the plate where the thickness  $h$  satisfies the condition

$$2h = n \frac{\lambda}{2},$$

the minima corresponding to odd values of  $n = 1, 3, 5, \dots$ , and the maxima to even values of  $n = 2, 4, \dots$ . Such conclusions can be drawn about transmitted light.

In the reflected light, the propagation difference for rays 1 and 2 at the normal incidence is  $AB + BC = 2h$ , i.e. is the same as for the transmitted light. It may appear that maxima and minima in the reflected light will be in the same regions of the plate as in the transmitted light. However, this would mean that the regions of the plate that reflect light the least transmit light the least as well. In particular, if the entire plate had the same thickness such that  $2h$



**Fig. 268.**  
Path of reflected and transmitted rays with the double reflection in a film.

<sup>2</sup> Both in transmitted and in reflected light, there are also rays undergoing multiple reflection. However, they are much weaker than the first two rays and can be neglected.

is equal to an odd number of half-waves, the plate would exhibit the smallest reflection and the minimum transmission. But since we assumed that the plate does not absorb light, the simultaneous attenuation of both reflection and transmitted light is impossible. It goes without saying that in a nonabsorbing plate, reflected light must supplement transmitted light so that dark regions in transmitted light correspond to light regions in reflected light, and vice versa. This conclusion is confirmed in experiments.

What is the error in our calculation of interference in reflected light waves? As a matter of fact, we ignored the difference in conditions of reflection. Some reflection acts occur at the interface between air and glass, while others take place at the glass-air interface (if we consider a thin glass plate in air). This difference leads to an additional phase shift corresponding to an additional propagation difference equal to  $\lambda/2$ . For this reason, the total propagation difference for rays reflected from the upper and lower surfaces of the plate of thickness  $h$  is  $2h + \lambda/2$ . The regions of minima correspond to the condition

$$2h + \frac{\lambda}{2} = m \frac{\lambda}{2},$$

where  $m$  is an odd number. The regions of maxima correspond to even values of  $m$ . Consequently, maxima and minima are formed in the regions of the plate where the plate thickness  $h$  satisfies the condition

$$2h = (m - 1) \frac{\lambda}{2} = n \frac{\lambda}{2},$$

where  $(m - 1)$  is denoted by  $n$ . Minima correspond to even values of  $n = 0, 2, 4, \dots$ , while maxima correspond to odd values of  $n = 1, 3, 5, \dots$ .

Let us compare the results obtained for determining the positions of maxima and minima in transmitted and reflected light. The positions of maxima and minima correspond to the film thickness determined from the condition  $2h = n(\lambda/2)$ , where

|                 | In transmitted light | In reflected light |
|-----------------|----------------------|--------------------|
| For an even $n$ | maximum              | minimum            |
| For an odd $n$  | minimum              | maximum            |

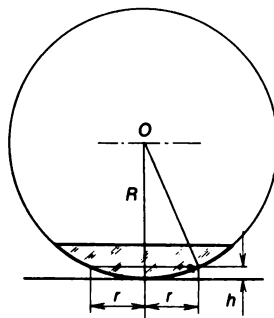
Thus, the regions of maxima in transmitted light correspond to the region of minima in reflected light and vice versa, in accord with experiments and considerations formulated above.

For Newton's rings, which are normally observed in reflected light (see Sec. 13.4), we find that the regions of maxima correspond to odd values of  $n = 1, 3, 5, \dots$ , while the regions of minima correspond to even  $n = 0, 2, 4, \dots$ . The central (zero) minimum corresponds to  $n = 0$  and has the form of a dark circle. The next first dark ring corresponds to  $n = 2$ , the second dark ring, to  $n = 4$ , and so on. Generally, the number  $N$  of a dark ring is connected to the number  $n$  through the relation  $N = n/2$ . The number  $N$  of a light ring is expressed in terms of  $n$  through the formula  $N = (n + 1)/2$ .

Instead of determining the thickness  $h$  of the region of the air film corresponding to a ring with a number  $N$ , it is more convenient to measure the diameter or radius of the corresponding ring. Figure 269 shows that  $R^2 = (R - h)^2 + r^2$ , and hence the thickness  $h$  of the air film is connected to the radius  $r$  of the ring and the radius  $R$  of the lens through the relation

$$(2R - h)h = r^2.$$

In experiments with Newton's rings, lenses with a very large radius  $R$  (of the order of several

**Fig. 269.**

To the calculation of the radii of Newton's rings.

meters) are used. Therefore, we can neglect the value of  $h$  in comparison with  $2R$  and simplify the last expression as follows:

$$2Rh = r^2, \quad \text{or} \quad 2h = r^2/R.$$

Thus, we obtain the following formula for determining wavelength  $\lambda$  with the help of Newton's rings:

$$2h = \frac{r^2}{R} = n \frac{\lambda}{2}.$$

If the radii of dark rings are measured, the number of the ring  $N = n/2$ . In this case, the wavelength is expressed by the formula

$$\lambda = \frac{r_N^2}{NR},$$

where  $r_N$  is the radius of the  $N$ th dark ring.

While measuring the radii of light rings, it should be borne in mind that  $N = (n + 1)/2$ . Accordingly, we obtain the following relation:

$$\lambda = \frac{2r_N^2}{(2N - 1)R},$$

where  $r_N$  is the radius of the  $N$ th light ring.

## Chapter 14

# Diffraction of Light

### 14.1. Bundles of Rays and the Shape of Wave Surface

For a very large number of problems in which the concepts of geometrical optics can be successfully applied, propagation of light was characterized with the help of rays. Then a parallel bundle of rays indicates that light energy propagates only in the direction of this bundle without dissipating in all directions so that the illuminance of a surface on which this bundle is incident remains unchanged at any distance from the source. A diverging beam indicates that light is distributed over an increasing surface so that the illuminance decreases in inverse proportion to the squared distance from the point from which the rays diverge (beam's origin). On the contrary, a converging beam indicates an increasing illuminance as the point of beam convergence is approached. The role of optical systems in geometrical optics is reduced to the transformation of wave fronts.

From the point of view of wave concepts, propagation of light is the propagation of waves, the role of rays being played by the lines normal to the wave-front surface. The nature of light propagation is determined by the shape of the wave front (*wave surface*). For example, a parallel bundle of rays corresponds to a *plane wave* whose front has the shape of a plane moving parallel to itself. Beams converging at a point or diverging from a point correspond to *spherical wave surfaces* whose centres lie at points of convergence or divergence of rays. A change in the curvature of a wave front indicates the change in the angle of convergence of rays.

Thus, the passage of a wave through a system of lenses or mirrors is reduced to a change in its front (Fig. 270).

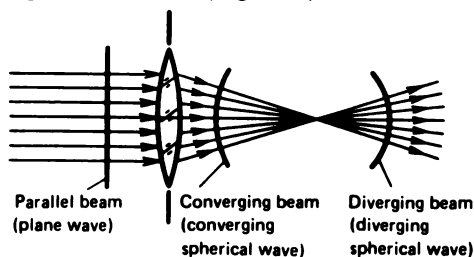


Fig. 270.

Change in the shape of a wave front as a result of the passage of the wave through a lens.

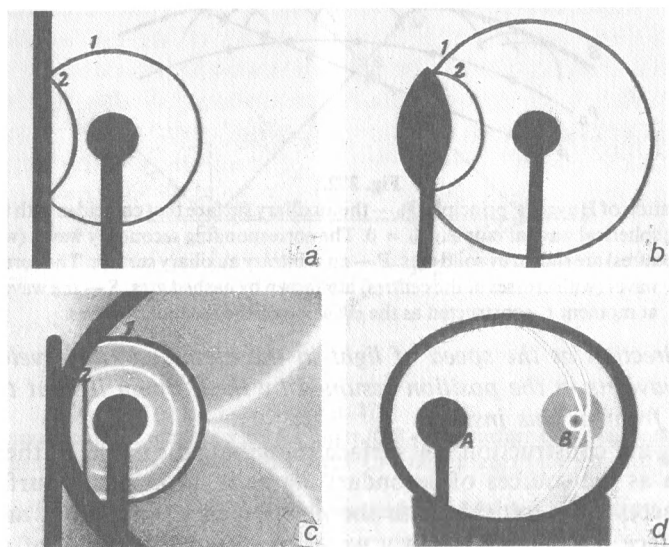


Fig. 271.

Change in the shape of the wave front upon reflection (the photographs of an acoustic wave in air): 1 — incident wave, 2 — reflected wave). Reflection of a spherical wave from (a) a plane mirror, during which the curvature of the wave front remains unchanged, (b) a convex mirror, during which the curvature of the wave front increases, (c) a parabolic mirror (the source is at the focus of the mirror), and the wave front becomes flat, (d) an elliptical mirror (the source is at focus A), as a result of which the wave converges at focus B.

The effect of reflection from different surfaces on the shape of a wave front is illustrated by photographs shown in Fig. 271 and representing the reflection of an acoustic pulse in air. Similar patterns can easily be obtained for refracted waves.

## 14.2. Huygens' Principle

The figures contained in the previous section give only a general qualitative idea about the wave nature of propagating light and about the effect of reflection and refraction on a light wave.

However, even Huygens in his time managed to use the idea of wave propagation in a medium for quantitative expression of the laws of reflection and refraction. For this purpose, he formulated a general principle observed in propagation of waves. The *Huygens principle* is the rule which makes it possible to determine the position of a wave front for a certain instant of time from its position at a previous close instant.

According to the Huygens principle, *every point of wave front may be regarded as a source of secondary spherical waves which spread out in the*

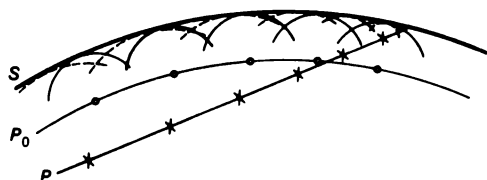


Fig. 272.

To the explanation of Huygens' principle.  $P_0$  — the auxiliary surface that coincides with the front of a diverging spherical wave at moment  $t_0 = 0$ . The corresponding secondary waves (with light circles at the centres) are shown by solid arcs.  $P$  — an arbitrary auxiliary surface. The corresponding secondary waves (with crosses at the centres) are shown by dashed arcs.  $S$  — the wave surface at moment  $t$ , constructed as the envelope of the secondary waves.

*forward direction as the speed of light in the medium. The envelope of all these wavelets in the position assumed by them by an instant  $t$  is the new wave front at this instant.*

In Huygens' construction, the surface containing the points of the medium chosen as the sources of secondary waves is an auxiliary surface. It does not necessarily coincide with the position of a wave front and can be the surface reached by primary waves at different instants of time.

In order to determine the position of the wave front by the moment of time  $t$ , we must construct the positions of secondary waves by this moment and draw an envelope of them. Thus, secondary waves propagating from the points reached by the primary wave at an earlier instant of time have time to spread out over longer distances, while secondary waves propagating from points which are taken as the sources later spread out over shorter distances.

Huygens' principle makes it possible to determine the required envelope by choosing an auxiliary surface in different ways, the ultimate result naturally being the same. The propagation of a spherical diverging wave whose front occupies position  $P_0$  at a certain instant  $t_0$  is analyzed in Fig. 272. Light from the source reaches different points of an auxiliary surface  $P$  at different instants of time. Thus, the application of Huygens' principle allows us to choose the centres of secondary waves in the most convenient way for solving a given problem. For this reason, Huygens' principle is successfully employed for analyzing various problems concerning the propagation of waves. The following section gives an example of application of Huygens' principle.

### 14.3. Reflection and Refraction from the Viewpoint of Huygens' Principle

Let a parallel bundle of rays fall on the interface  $ab$  between two media (Fig. 273), forming angle  $i$  with the normal to the interface. According to the reflection rule, the bundle of refracted rays will propagate in the direc-

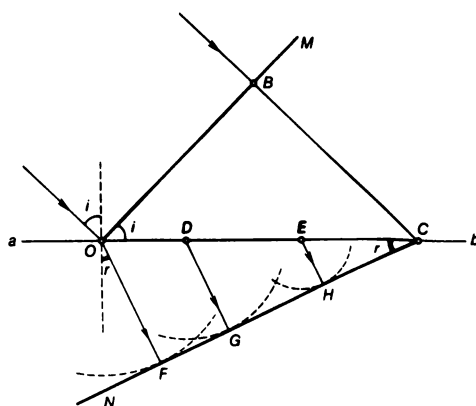


Fig. 273.

To the derivation of the refraction law for waves.  $OB$  — the surface of the incident wave,  $ab$  — the interface between two media, and  $NC$  — the surface of the refracted wave.

tion specified by angle  $r$ . The experimentally derived reflection law states that

$$\frac{\sin i}{\sin r} = n,$$

where  $n$  is the refractive index of the second medium relative to the first one. This quantity is independent of the angle of incidence  $i$  and characterizes the properties of the two media.

According to the wave theory, the problem can be formulated as follows. A plane wave whose surface forms an angle  $i$  with the interface is incident on it. The velocities of wave propagation in the first and second media are  $v_1$  and  $v_2$  respectively.

For deriving the law of refraction and determining the refractive index, we shall make use of Huygens' principle. The problem can easily be solved if we take for the centres of secondary waves points lying on the interface. Let us suppose that the incident plane wave reaches the interface at the instant  $t = 0$  at point  $O$ , i.e. the incident wave surface occupies position  $OM$ . Let us determine the position of the envelope at instant  $t = \tau$  when point  $B$  of the surface of the incident wave reaches the interface at point  $C$ . Since the velocity of the wave in the first medium is  $v_1$ , distance  $BC$  is equal to  $v_1\tau$ . The secondary wave emerging from point  $O$  propagates in the second medium during this time over distance  $OF = v_2\tau$ . The primary wave will reach point  $D$  somewhat later, and the secondary wave from this point will penetrate into the second medium by the moment  $\tau$  to a smaller depth equal to  $DG$ . The penetration depth  $EH$  of the secondary wave originating at point  $E$  will be still smaller. By the moment  $\tau$ , no wave will start propagating from point  $C$  since the primary wave just reaches



point  $C$  by this moment. Having constructed the envelope (which turns out to be a plane tangent to all secondary spherical waves), we shall find line  $CN$ , viz. the position of the front of the refracted wave. This wave front propagates in the second medium at a velocity  $v_2$  along  $OF$  ( $\perp CN$ ) determined by angle  $r$ .

The relation between angles  $i$  and  $r$ , i.e. the *law of refraction*, can be determined from  $\triangle OBC$  and  $\triangle COF$ . Indeed,  $BC = v_1\tau = OC \sin i$  and  $OF = v_2\tau = OC \sin r$ , whence

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}.$$

If we denote the ratio  $v_1/v_2$  by  $n$ , we shall obtain the law of refraction in the conventional form  $\sin i/\sin r = n$ . The value of  $n$  does not depend on angles  $i$  and  $r$  and is known as the *refractive index*.

Thus, we have not only derived, by using Huygens' arguments, the correct law of refraction but also clarified the physical meaning of the refractive index  $n$ : *the refractive index is equal to the ratio of the velocities of a light wave in the first and second media*.

If the first medium is air (or vacuum which is practically the same for many problems), and the second medium is water, it is known from experiments that  $n = 1.33$ . Thus, our line of reasoning leads to the conclusion that the velocity of light in air (vacuum) is 1.33 times higher than in water. It will be shown (see Sec. 17.5) that direct measurements of the velocity of light in water and in air confirm this conclusion.

Similar technique can be used for analyzing the reflection of a wave. The result will be as follows: *the angle of reflection is equal to the angle of incidence*.

#### 14.4. Huygens' Principle in Fresnel Interpretation

The previous section demonstrated that Huygens' principle is fruitful for solving many important optical problems. In Huygens' formulation, this principle had the form of a geometrical rule according to which the result of action of secondary waves can be determined by constructing the surface which envelopes these waves. The French physicist Augustin Fresnel (1788-1827) borrowed from Huygens' principle the idea about secondary waves and applied the laws of interference to them. According to Fresnel, *the rule for constructing the envelope should be replaced by the calculation of the mutual interference of secondary waves*. This calculation led to the same results as those provided by Huygens' rule in its initial form.

Fresnel's method not only provides a deeper physical meaning to Huygens' principle but also makes it possible to solve new problems which could not be analyzed by using the initial rule formulated by Huygens.

Let, for example, a wave propagate in a certain direction in a homogeneous medium. Any point reached by the wave becomes a source of secondary waves propagating in all directions. It may appear that the initial direction of propagation will change as a result and the light wave will dissipate in all directions. If, however, following Fresnel we take into account the mutual interference of these secondary waves, it will turn out that secondary waves suppress one another in lateral directions and enhance one another only in the initial direction. For this reason, light propagates only in the initial direction. Consequently, we arrive at the explanation of the rectilinear propagation of light in a homogeneous medium.

If, however, the medium is not homogeneous (e.g., contains exogenous inclusions or consists of different media as in the case when mirrors, plates or lenses are arranged in air), the result will be different. Passing through such a medium, light does not propagate in a straight line but is scattered in different directions, undergoes reflection, refraction, etc. It was shown, for example (see Sec. 14.3), how Huygens' principle can be used for deriving quantitative laws of refraction and reflection.

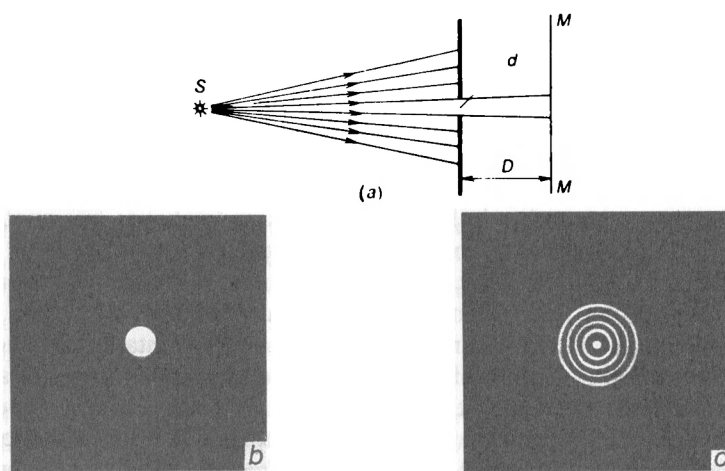
Thus, the basic laws of geometrical optics, viz. the law of rectilinear propagation of light, and the laws of reflection and refraction, can be interpreted from the viewpoint of the wave theory with the help of the Huygens-Fresnel principle.

It is even more important that this principle can be used for analyzing optical phenomena under the conditions when the laws of geometrical optics cease to be valid.

#### 14.5. Simple Diffraction Phenomena

The simplest case of violation of the laws of geometrical optics was described in Sec. 9.2 where it was shown that the law of rectilinear propagation of light is not observed when light passes through a very small aperture: at the edges of the hole light bends into the shadow region and does not propagate in a straight line. Such a bending can be observed in any case when a shadow from an object is cast on a screen even when the object is not very small. But since the angle of deviation from rectilinear direction of propagation is usually small, the effect becomes stronger if the screen is arranged far from the object.

For example, according to the rules of geometrical optics, light emitted by a small bright source through a circular aperture of diameter  $d$  (Fig. 274a) must produce on screen  $MM$  a sharp light circle against a dark background (Fig. 274b). This pattern is actually observed under the normal conditions of the experiment. If, however, the distance between the hole and the screen is a few thousand times longer than the diameter of the hole, important details of the phenomena become noticeable: a more com-

**Fig. 274.**

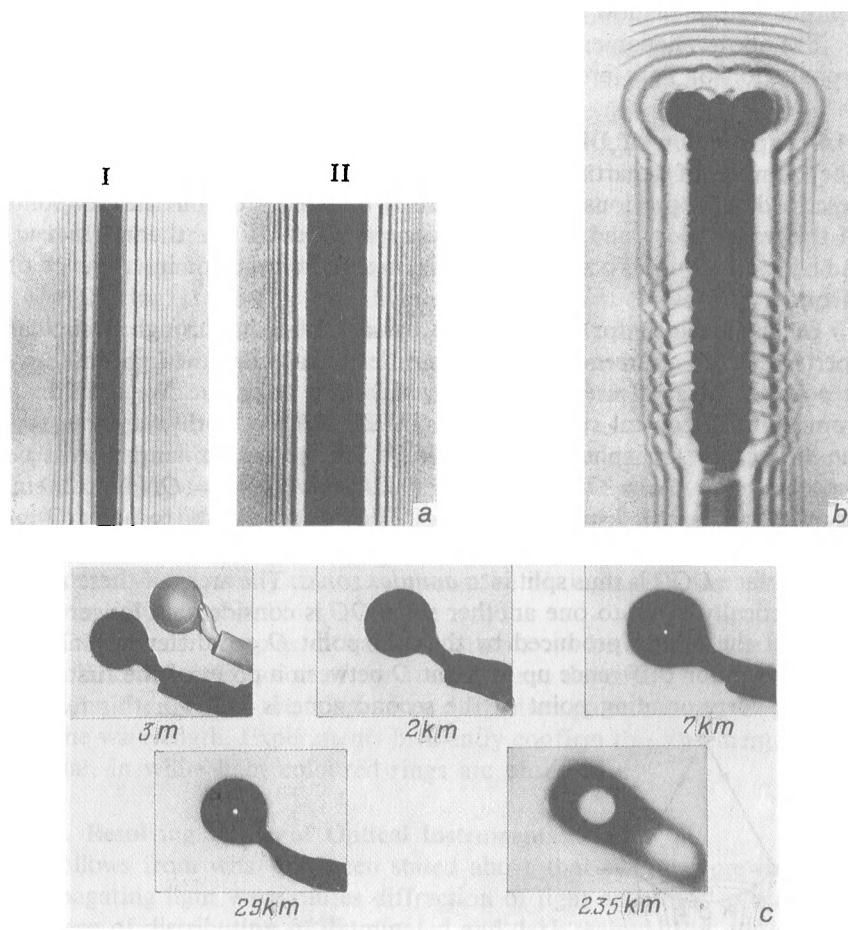
Diffraction by circular aperture: (a) schematic diagram of the experiment, (b) the shape of the shadow when the aperture diameter  $d$  is comparable with the distance  $D$  from the aperture to the screen, (c) the shape of the shadow when the aperture diameter  $d$  is smaller than the distance  $D$  from the aperture to the screen by a factor of  $10^3$ .

plex pattern is formed, consisting of a system of light and dark concentric rings gradually transforming into one another (Fig. 274c). For another ratio between the hole diameter and the distance to the screen, a dark spot may appear at the centre of the pattern. This experiment gives a visual idea about the wave properties of light and cannot be explained in any way from the point of view of geometrical optics (for details, see Sec. 14.6).

Thus, for observing diffraction in the experiment described above, one should either use a very small aperture (for a laboratory experiment, of the order of hundredths of a millimetre) or arrange a screen at a large distance from the aperture (of the order of hundreds of metres if a hole having a size of a few millimetres is used).

Similarly, illuminating sufficiently large opaque objects located at a comparatively short distance from a screen by a small source, we obtain quite a sharp shadow. If, however, the separation between the object and the screen is considerably larger than the size of the object, the shadow has an intricate form.

Figure 275a shows the shadow of a rectilinear object (pencil or wire) cast on a remote screen. There are regions in the shadow into which light penetrates, and the edges of the shadow are bordered by a number of light and dark fringes. Figure 275b shows the shadow from a screw, obtained under similar conditions. The intricate nature of the pattern indicates that

**Fig. 275.**

Photographs of diffraction patterns (the shadow is cast on the screen): (a) diffraction by a wire (I) and a pencil (II), (b) diffraction by a screw, (c) diffraction by a hand keeping a plate for different distances from the hand to the screen.

light deviates considerably from straight lines and bends about the edges, producing a system of light and dark regions which hardly resemble a sharp shadow similar to the object. Figure 275c shows the shadow cast by a hand with a plate. The experiments were carried out in 1912 by V.K. Arkad'ev and A.G. Kalashnikov at the Moscow University with a reduced model of the hand with a plate in it. The distances from the model to the screen, indicated in the figure, are recalculated for the experiment with a plate of

a real size. The farther is the screen, the less the resemblance between the contour to the shadow and the object.<sup>1</sup>

The above phenomena involving the violation of the law of rectilinear propagation of light are known as *diffraction of light*.

#### 14.6. Explanation of Diffraction by Fresnel's Method

The examples of departure from the law of rectilinear propagation of light described in the previous section can easily be explained from the viewpoint of the wave theory and are natural consequences of this theory. Indeed, the light distribution observed in each case is the result of interference of secondary waves.

Let us consider, for example, the passage of light through a circular aperture  $DD$  in a screen (Fig. 276). In order to calculate the light intensity at point  $O$ , we shall use the following auxiliary technique. We shall draw from point  $O$  conical surfaces  $OKL$ ,  $OMN$ ,  $OPQ$ , etc. till they intersect the surface of the spherical wave  $DCD$ . The generatrix lengths will be chosen so that  $OL = OC + \lambda/2$ ,  $ON = OL + \lambda/2$ ,  $OQ = ON + \lambda/2$ , etc. In other words, the distances from points  $C$ ,  $L$ ,  $N$ ,  $Q$ , ... to point  $O$  increase by half a wavelength ( $\lambda/2$ ) of the light incident on the hole. The wave surface  $DCD$  is thus split into *annular zones*. The areas of these zones are practically equal to one another since  $OC$  is considerably longer than  $\lambda/2$ . But the effects produced by them at point  $O$  are different. Indeed, the propagation difference up to point  $O$  between a point of the first zone and the corresponding point of the second zone is  $\lambda/2$ . For this reason,

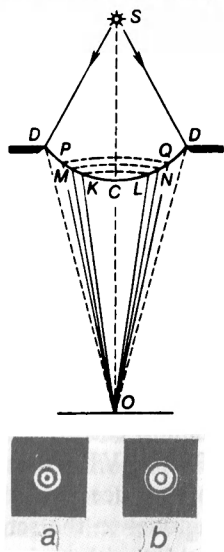


Fig. 276.

To the explanation of diffraction by a circular aperture. At the bottom, the schematic representation of the observed pattern: (a) for an odd number of zones, (b) for an even number of zones.

<sup>1</sup> These photographs were taken by V. K. Arkad'ev and are borrowed from his paper.

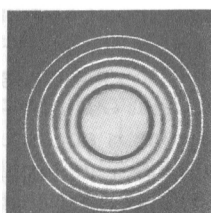
light waves from the first and second zones, having reached point  $O$ , will suppress each other so that the effect of the first zone at this point is practically cancelled by the action of the second zone. A similar line of reasoning leads to the conclusion that the effect of the third zone at point  $O$  is opposite to that of the second zone, the action of the fourth zone is opposite to the action of the third zone, and so on. Generally, any two neighbouring zones practically cancel effects. If the diameter  $DD$  of the hole is such that it incorporates just two zones, point  $O$  will almost be unilluminated since two adjacent zones mutually suppress each other. Light will be mainly distributed around point  $O$  so that we shall see a dark spot surrounded by a bright ring. If the hole contains three zones, point  $O$  will be bright since the third zone suppresses the action of the second zone, and the point will be illuminated due to an almost unsuppressed action of the first zone. Then the bright central point will be bordered with a dark ring followed by a bright region. In general, if the number of zones is even, we have a dark spot at the centre, surrounded by alternating light and dark rings. For an odd number of zones, we have a bright spot at the centre, surrounded by dark and light rings. The size of these rings is the smaller, the larger the hole diameter. When a diameter is large, dark and light rings near the centre alternate so frequently that they cannot be distinguished, and diffraction pattern is blurred.

Other, more complex diffraction patterns can be explained in a similar way. Since the calculation of the Fresnel zones depends on the wavelength of light, the form of the fringes in a diffraction pattern is also determined by the wavelength. Experiments brilliantly confirm this statement. In particular, in white light coloured rings are observed.

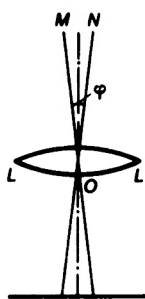
#### 14.7. Resolving Power of Optical Instruments

It follows from what has been stated above that the aperture limiting a propagating light wave causes diffraction of light and leads to a complex pattern of distribution of illuminated and dark regions. But every optical instrument, including the eye, is supplied with lenses or mirrors that always restrict a wave front. Thus, it should be expected that a diffraction pattern is always formed when an image is obtained with the help of an optical system.

Indeed, detailed calculations and accurate experiments show that the image of bright point formed by an objective is not just a bright point against a dark background but rather a complex pattern of dark and light rings gradually transforming into one another and merging with the surrounding dark background (Fig. 277). The larger the diameter of the objective forming the image, the finer the diffraction pattern, i.e. the closer are diffraction rings in it. Normally we ignore this complication and assume that the image of a bright point is just a light point. However, this complica-

**Fig. 277.**

The image of a luminous disc (say, a planet) obtained with the help of a telescope (diffraction pattern).

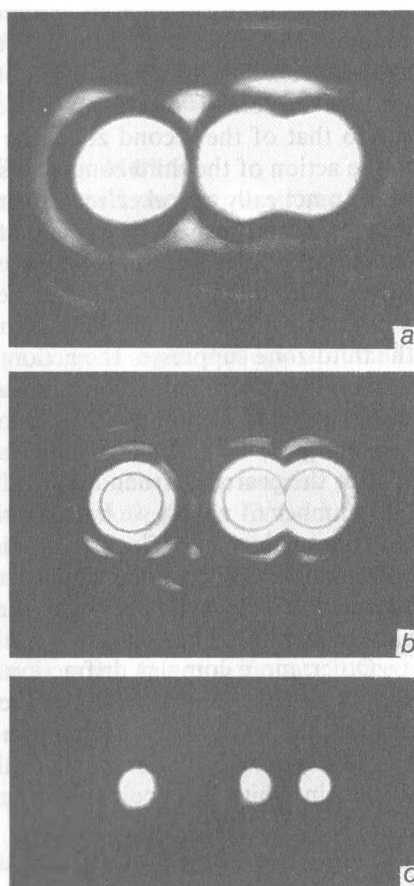
**Fig. 279.**

To the explanation of the resolving power of a telescope:  $OM$  and  $ON$  — directions to two close stars,  $\varphi$  — the angular separation between the stars, and  $LL$  — the telescope objective. At the bottom, the schematic of a negative image.

tion always takes place and can be revealed in more thorough observations. It cannot be eliminated with any design of the objective since it is due to the wave nature of light.

It is interesting to note that the degree of diffraction distortion is reduced with increasing diameter of the objective (Fig. 278). On the contrary, the distortions due to an error of an objective such as spherical aberration are the stronger, the larger its diameter (see Sec. 11.6).

The defects of objectives of photographic cameras usually play a more significant role than distortions introduced by diffraction. For this reason, the reduction of the objective diameter (stopping down), which weakens

**Fig. 278.**

Decreasing diffraction distortions by increasing the diameter of the objective (bottom view).

the effect of these defects, normally improves the sharpness of the image. For very small aperture diameters, however, the distortion due to diffraction starts to prevail. The defects of perfect astronomical objectives are so small that distortions are mainly introduced by interference in spite of the fact that these objectives usually have large diameters (10 cm and larger).

Diffraction sets the limit for distinguishing the details of an object while using an optical instrument. Let us suppose, for example, that two stars with a small angular separation are viewed through a telescope (Fig. 279). If the telescope were perfect, according to the laws of ray optics we would have two closely spaced sharp point-like images. However, due to diffraction we obtain a pattern in the form of two systems of light and dark rings instead of two separate points (Fig. 279 bottom).

If the centres of these systems are located very close to each other (the stars are close in direction) and the rings are not very small (a small objective diameter), the images superimpose to form a pattern that differs insignificantly from the system of rings surrounding the image of a single star. It is impossible to establish the separate location of the stars with the help of this pattern: *the instrument is unable to resolve two such close stars*. Thus, the ability of an optical instrument to discern the details is limited by the wave nature of light. This property of objectives is known as the *resolving power*, or resolution. Objectives with a large diameter have a high resolving power. For example, a telescope with an objective diameter of 12.5 cm can resolve two stars with an angular separation of 1", while a half-metre telescope objective is capable of resolving two stars separated by 0.25". Thus, through a large telescope one can sometimes see individual close stars (star clusters) which merge into a single bright spot when viewed through a small telescope and cannot be distinguished from nebulae. This explains the tendency to construct telescopes with large diameters of objectives. The other reason for that was indicated in Sec. 12.10.

This limitation of the ability to discern details is also imposed on the human eye having a pupil diameter of about 2-4 mm. For this reason, the eye resolves bright points with an angular separation of about a minute.<sup>2</sup> Similar considerations set the limit on the resolving power of a microscope (Sec. 12.7), where the objective diameter also restricts the bundles of rays participating in the image formation.

The resolving power of an optical instrument should not be confused with its magnification (see Sec. 11.4). In an enlarged image obtained by using an optical instrument is viewed through another optical instrument, the magnification can be made as large as desired. However, this does not

---

<sup>2</sup> The resolving power of the eye, which is determined by the diameter of the pupil, also depends on the complex structure of the retina. This structure sets the limit on the resolving power of the eye, which is also about 1' (at a sufficiently high illuminance).



improve the resolving power of the system of instruments. Indeed, the image obtained with the help of the first instrument will contain only details that can be resolved by it. A further magnification of this image (which does not contain fine details) naturally cannot reconstruct them but can even blur some details of the first image. Consequently, the resolving power of the system of instruments cannot exceed the resolution of the worst of them.

### 14.8. Diffraction Grating

The position of maxima and minima forming a diffraction pattern depends, as was shown above, on the wavelength  $\lambda$  of light. For this reason, when diffraction is observed in a composite light (say, in white light), in which different wavelengths are present, diffraction maxima for different colours will be in different regions, i.e. the dispersion of white light is observed in the diffraction pattern.

The most interesting and practically important case of diffraction where the dispersion plays a leading role can be realized with the help of a *diffraction grating*.

A simple diffraction grating is a plate with alternating narrow transparent and opaque parallel strips. Such a grating can be obtained, for example, by ruling a number of blazes with a diamond glass cutter on a glass plate, so that narrow strips remain intact. High-quality gratings are prepared by ruling the surface of a metal mirror. In such gratings, the strips reflecting light and scratched regions scattering light in all directions alternate. These appliances are known as *reflection gratings*. The sum of the widths of a transparent (reflecting) and opaque (scattering) strips is called the *grating period*  $d$ . In the best of modern gratings, there are up to 1800 slits per millimetre so that the grating period is about  $0.8 \mu\text{m}$ .

Let us direct a parallel beam of rays on a grating at right angles to its surface. For this purpose, we can illuminate by bright light a narrow slit  $S$  arranged in the focal plane of a converging lens  $L_1$  (Fig. 280). Passing through narrow transparent slits of grating  $RR$ , light undergoes diffraction, deviating sideways from its initial direction. By using a second lens  $L_2$ , we obtain an image of slit  $S$  on screen  $M$ . Since as a result of diffraction the rays are incident on lens  $L_2$  at different angles, the images of slit  $S$  will be formed in different regions of the screen. However, as a result of mutual interference of diffracted rays, some of these images will not appear (minima), while others will be especially bright (maxima  $S_0, S_1, S'_1, S_2, S'_2, \dots$ ).

The result of this interference can be calculated by using Fig. 281, which shows a few adjacent opaque regions of the grating. Let us assume that a monochromatic light of wavelength  $\lambda$  is incident on the grating. Suppose that the front of the incident wave coincides with  $AB$  (the grating plane),

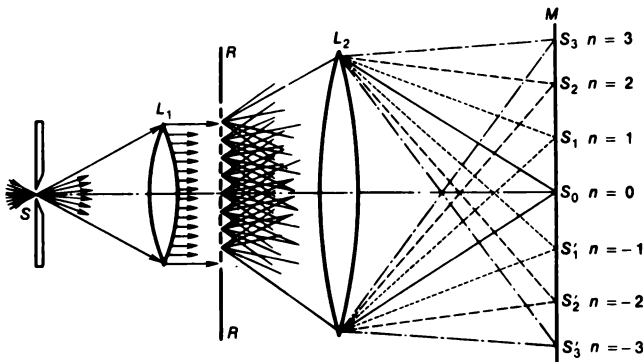


Fig. 280.

Operation of a diffraction grating:  $S$  — a brightly illuminated slit parallel to the grating lines,  $L_1$  — the lens in whose focal plane the slit is located,  $R$  — diffraction grating,  $L_2$  — the lens producing, together with  $L_1$ , the image of  $S$  on screen  $M$ ,  $S_0$  — the image of slit  $S$  formed by nondiffracted rays (the zero-order maximum),  $S_1$  and  $S_1'$  — the images of slit  $S$  formed by the rays diffracted by the grating (the first-order maxima),  $S_2$  and  $S_2'$  — the images of slit  $S$  formed by the rays diffracted by the grating (the second-order maxima), etc.

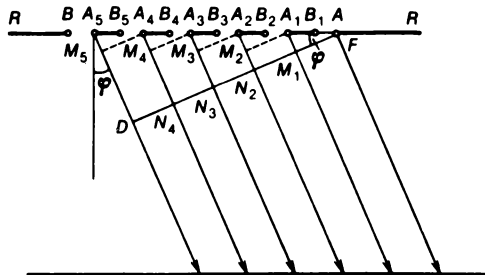


Fig. 281.

To the theory of diffraction grating.

i.e. light is incident at right angles to the grating. As a result of diffraction of light, at the grating outlet we shall observe light waves propagating in different directions. Let us consider the waves propagating from the grating in a direction forming angle  $\varphi$  with the normal to the grating plane. The path differences for the rays emerging from similar points of slits, say, from the right edges (points  $A, A_1, A_2, A_3, \dots$ ), from the left edges ( $B_1, B_2, B_3, B_4, \dots$ ) or from the middles of the slits naturally have the same value. These differences are

$$\begin{aligned} A_1M_1 &= AA_1 \sin \varphi = d \sin \varphi, \\ A_2M_2 &= A_2N_2 - A_1M_1 = 2d \sin \varphi - d \sin \varphi = d \sin \varphi, \\ A_3M_3 &= A_3N_3 - A_2N_2 = 3d \sin \varphi - 2d \sin \varphi = d \sin \varphi, \text{ and so on,} \end{aligned}$$

where  $d = AA_1 = A_1A_2 = A_2A_3$  is the period of the grating. In order that all beams enhance one another, it is necessary that  $d \sin \varphi$  be equal to an

integral number of wavelengths  $\lambda$ , i.e.

$$d \sin \varphi = n\lambda, \quad (14.8.1)$$

where  $n$  is an integer. Thus, condition (14.8.1) is the condition of mutual amplification of all the beams passing through the grating slits. This condition allows us to determine the values of angle  $\varphi$ , i.e. the directions in which the maxima of light intensity of wavelength  $\lambda$  will be observed. These angles can be found from the formula

$$\sin \varphi = n\lambda/d \quad (14.8.2)$$

by ascribing different integral values to  $n$ : 0,  $\pm 1$ ,  $\pm 2$ ,  $\pm 3$ , etc.

### 14.9. Diffraction Grating as a Spectral Instrument

Formula (14.8.2) shows that several maxima can be observed for a given wavelength  $\lambda$ . The direction corresponding to  $n = 0$  is characterized by  $\varphi = 0$ . This is the direction of the initial beam. The corresponding maximum is known as the *zero-order maximum*. Point  $S_0$  in Fig. 280 corresponds to it. For  $n = 1$ , we have  $\sin \varphi_1 = \lambda/d$ , for  $n = -1$ ,  $\sin \varphi'_1 = -\lambda/d$ , i.e. we have two first-order maxima arranged symmetrically on both sides of the zero-order maximum (points  $S_1$  and  $S'_1$  in Fig. 280). For  $n = \pm 2$ , we obtain  $\sin \varphi_2 = 2\lambda/d$  and  $\sin \varphi'_2 = -2\lambda/d$ , i.e. two symmetrical second-order maxima (points  $S_2$  and  $S'_2$  in Fig. 280), and so on.

Hence it immediately follows that for waves having different wavelengths  $\lambda$  the positions of the zero-order maxima corresponding to  $\varphi = 0$  coincide, while the positions of the maxima of the first, second, etc. order are different: the larger  $\lambda$ , the higher the corresponding values of  $\varphi$ . Thus, longer waves produce the image of the slit at a larger distance from the zero-order maximum. If a composite (say, white) light is incident on slit  $S$  (Fig. 280), we obtain a number of coloured images of the slit, arranged in the order of increasing wavelengths. In the region of the zero-order maximum which is the same for all wavelengths, we obtain the white image of the slit, coloured fringes being arranged on both sides of it from violet to red (first-order spectra). The next sets of coloured fringes will be arranged somewhat farther (second-order spectra), etc.

Since the wavelength of red light is about 760 nm, while that of violet light is about 400 nm, the red region of the second-order spectrum overlaps with the third-order spectrum. Higher-order spectra overlap even more. Figure V (see end leaf) schematically represents a spectrum obtained with the help of a diffraction grating. It can easily be seen that this experimental-ly obtained diagram confirms the conclusions drawn earlier.

If the period  $d$  of a grating is small, the corresponding values of  $\varphi$  are large. Similarly, for a small  $d$  the difference between two values of  $\varphi$  cor-

responding to different wavelengths is large. Thus, by reducing the grating period, we increase the angular separation of the maxima corresponding to different wavelengths. If a light incident on a slit is a combination of different wavelengths  $\lambda_1, \lambda_2, \lambda_3, \dots$ , using a diffraction grating it is possible to separate more or less completely these wavelengths. The larger the grating size, i.e. the larger number of slits it incorporates, the higher its quality: with an increased number of slits, more light passes through the grating (maxima become brighter), and radiations at close wavelengths are separated better (maxima become sharper).

A grating with a known period can be used for determining the wavelength by measuring angle  $\varphi$  which determines the position of the maximum of a given order. In this case, the relation  $d \sin \varphi = n\lambda$  gives

$$\lambda = \frac{d \sin \varphi}{n}. \quad (14.9.1)$$

The values of wavelength measured with the help of diffraction gratings are among the most accurate.

### 14.10. Preparation of Diffraction Gratings

A high-quality diffraction grating must have a small period and a large number of slits. In modern high-quality gratings, this number exceeds 100 000 (the grating width is up to 100 mm, the number of slits is 1200 per millimetre). The slits must be strictly parallel to one another, and the widths of each type of slits (transparent and opaque) must be strictly identical (the equality of the widths of transparent and opaque slits is not required). It is essential that the period  $d$  of a grating be constant.

High-quality gratings are prepared by ruling parallel lines on the surface of a metal mirror (reflecting grating), with the help of a sharp stylus, the slits scattering light in all directions playing the role of dark slits, while intact regions of the mirror, the role of light slits. For preparing a transmission grating, a glass plate can be ruled.<sup>3</sup> A first-class ruling machine is required for obtaining gratings. At present, diffraction gratings produced by recording the interference pattern formed as a result of interference of two plane monochromatic waves incident on a photographic plate at different angles have found a wide application.

### 14.11. Diffraction at an Oblique Incidence of Light on a Grating

Figure 280 shows the diffraction of a parallel beam of rays (plane wave) in the case when the incident beam is normal to the plane of the grating (the angle of incidence is zero). Naturally, diffraction will also be observed with an oblique incidence of light when the angle of incidence is  $\alpha$ .

In this case, diffraction occurs as if we replaced the grating by another one, representing its projection on the direction normal to the incident rays (Fig. 282). The zero-order maximum will consequently lie on the continuation of the primary beam, and the period will be

<sup>3</sup> Since the stylus is blunted as a result of ruling slits on the surface of glass or metal, it is difficult to ensure equal widths of the slits, and hence high-quality gratings are seldom prepared with glass. Transmission gratings are made in the form of replicas from special plastic materials from a metal (reflection) grating. Such gratings (*replicas*) are comparatively cheap.

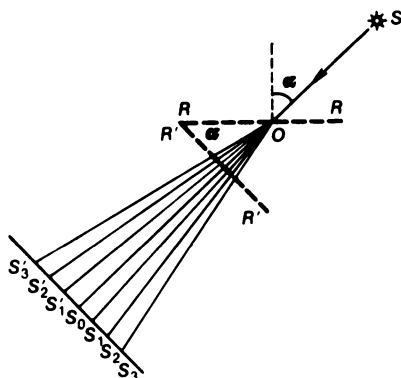


Fig. 282.

Schematic representation of diffraction for an oblique incidence of a light beam on a grating:  $SO$  — the direction of the primary beam,  $\alpha$  — the angle of incidence,  $RR$  — the diffraction grating,  $R'R'$  — the projection of  $RR$  on the direction perpendicular to the primary beam,  $OS_0$  — the direction to the zero-order maximum,  $OS_1$  and  $OS_1'$  — the directions to the first-order maxima,  $OS_2$  and  $OS_2'$  — the directions to the second-order maxima, etc.

$d' = d \cos \alpha$ . When  $\alpha$  is close to  $90^\circ$  (sliding incidence), the period determining the diffraction pattern may be much smaller than the period of the actual grating. Therefore, diffraction of light can be observed with the help of a very rough grating.

Taking, for example, a metal rule with millimetre divisions and placing it obliquely to the rays emitted by the filament of a remote incandescent lamp (the filament must be arranged parallel to the grating slits to play the role of an illuminated slit), we can easily observe diffraction spectra of various orders. By varying the angle of incidence (rotating the rule), we can observe the extension of spectra and an increase in the separations between orders (i.e. a reduction in the grating period) as the angle of incidence approaches  $90^\circ$ .

Using oblique incidence, it is possible to observe the diffraction of X-rays with the help of an ordinary diffraction grating although the wavelength of X-rays is about  $10^{-4}$  that of ordinary light. For example, by placing a grating having a period of  $1 \mu\text{m}$  at an angle  $\alpha = 89^\circ 59' 40''$ , we obtain a pattern corresponding to a grating with a period of about  $1 \text{ \AA}$ , which can be used for investigating the diffraction of X-rays having a wavelength of a fraction of an angstrom.<sup>4</sup> This method of observation made it possible to measure the wavelength of X-rays quite accurately.

<sup>4</sup> It should be recalled that  $1 \text{ \AA} = 10^{-10} \text{ m} = 0.1 \text{ nm}$ .

## Chapter 15

# Physical Principles of Optical Holography

In this chapter, we shall consider a rapidly developing method of obtaining *three-dimensional* images of various objects, which has at present a large number of scientific and technical applications. This method is known as optical holography and is based on interference and diffraction of light.

### 15.1. Photography and Holography

In order to obtain a photograph of a nonluminous object, it is illuminated, and its real image is formed with the help of an optical system (objective or spherical mirror) on a photographic plate (film) which is then developed and fixed.

The photographic technique has reached a high level and become extremely important for science and engineering. It will undoubtedly remain of great value in future as excellent and simple means of recording valuable information accessible to optical methods of observation.

In spite of a high level of development of instrumental optics and photographic technique, however, the potentialities of photography are limited in some respects. Let us briefly analyze the restrictions inherent in this traditional method of recording optical information.

1. An optical system is required for obtaining the image of an object on a screen or photographic plate.

2. The optical system forms the image of a three-dimensional object on a plane screen or photographic plate. Under optimal conditions, only those points of the object that lie in a certain plane normal to the optical axis of the system participate in image formation.

3. The image obtained on a screen or photographic plate does not allow one to view the object from different directions as in the case when it is observed directly. In other words, the third dimension of the object is lost in photorecording.

4. Each region of the surface of a photographic plate contains information only about a certain detail of the object. For this reason, having a part of a negative, it is impossible to observe the complete image of the object.

5. It is unexpedient to record the images of a few objects on the same negative if the images overlap: the information about one object deteriorates the perception of information about another object.

Let us now consider from a more general viewpoint the extent to which a photograph makes it possible to use the information about the object, which is transported by the electromagnetic field that is reflected by the object.

The optical image on a screen or photographic plate is produced due to nonuniform illumination of their surfaces by the light reflected by the object. The illuminance is measured by the light energy incident per unit time on a unit area. Since the frequency of optical electromagnetic radiation is high, the light energy is determined by the time-averaged value of the energy flux.

In turn, the average value of the energy flux depends on the amplitude of the electric vector  $\mathbf{E}$  and the magnetic vector  $\mathbf{B}$  of the light field near the surface of every region of the image and does not depend on the initial phase of the field oscillations in a given region. For example, two regions of an image will be equally illuminated if the amplitudes of vectors  $\mathbf{E}$  and  $\mathbf{B}$  near these regions are equal to each other respectively even if their phases are different.

Obviously, the photographic recording of illuminance distribution in the image plane does not allow us to take into account the distribution of oscillation phases in this plane. Indeed, the darkening of a photographic negative is only due to the energy absorbed by it, and the energy depends on the illuminance of the negative and on the exposure time.

Before going over to the analysis of the fundamentals of holography, we shall explain some terms that will be used in the further analysis. A light wave is termed *monochromatic* if it contains the radiation of a strictly definite wavelength. Real light sources naturally have no such property, but if the wavelength interval in radiation is small, the wave will also be treated as monochromatic. If the phase difference for two waves arriving at a point of space does not change with time, these waves exhibit *time coherence* and are capable of forming a stable interference pattern.

A light beam is known to be *spatially coherent* if the phase difference at two points of a plane normal to the direction of its propagation remains constant.

If an object of observation is illuminated by a nonmonochromatic and spatially incoherent light, the phases of waves reflected by the object are distributed in the image plane at random (in space and time) and can give no additional information about the object.

A different situation is observed when an object of observation is illuminated by a monochromatic and spatially coherent light beam. In this case,

the phase distribution of light waves reflected by the object is governed by definite laws and contains information on the object, which supplements the information provided by the amplitudes of waves.

For example, the phases of the waves reflected by far regions of the object will lag behind and have a different distribution in the image plane in comparison with the phases of the waves reflected by points lying close to the optical system. Consequently, the phase difference for waves reflected by a three-dimensional object may contain information about the extension of the object along the direction of observation. However, as was mentioned above, the photographic method of recording images provides no opportunity for using the phase information. New methods are required for decoding this information.

A recently developed branch of optics, viz. holography, solves the problem of a more complete utilization and recording of information carried by the field of light waves reflected by an object. The generally adopted name of the new trend in optics meaning "*complete recording*" (of light field) (from the Greek *holos* for "the whole" and *graphein* for "to write") fully reflects the goal aimed by the founder of holography, the English physicist Dennis Gabor.

The first stage of a holographic recording of optical information is the registering of both amplitude and phase characteristics of the wave field reflected by an object of observation. Under certain special conditions which will be specified in detail later, this recording is carried out by photographic methods but without forming an optical image of the object. The photographic plate with such a special record of the field parameters is known as a *hologram*.

The next stage of holographing is the extraction from the hologram of the information about the object, which is recorded on it. For this purpose, the hologram is illuminated by a light beam (in some cases, the light reflected by a hologram is used).

A hologram is a peculiar two-dimensional (sometimes three-dimensional) structure on which the incident light undergoes diffraction. Having been diffracted at the hologram, the light beam may form on a screen a real optical image of the object without using any optical system. This beam can also produce a wave field equivalent to the one propagating earlier (during hologram recording) from the object of observation. In order to use this wave field for obtaining information about the object of observation, an optical system is required.

A remarkable property of a hologram, which reflects its destination (complete recording), consists in a large amount of information registered in it.

A hologram makes it possible to completely reconstruct in the absence



of an object the wave field produced earlier (during the hologram recording) by the object itself. Using this field, it is possible to obtain not only one but a large number of images as during a direct observation of the object from different points of view. This is the most significant distinction of a hologram from a photograph.

Using the holographic method, it is possible, for example, to reconstruct the effect of the three-dimensional nature of an object (i.e., observe a *parallactic displacement*<sup>1</sup> with a change in the position of an observer) and the colours of the object surface without using conventional methods of colour photography.

The utilization of the wave field of an object reconstructed during the illumination of a hologram for obtaining the optical information about the object explains the other name attached to this method, viz. image formation by reconstructing its wave field.

### 15.2. Holographic Recording with a Plane Reference Wave

It was mentioned above that it is impossible to record directly the phase of optical oscillations by the methods which register only the time-averaged intensity of light. However, it is known that in interference phenomena the light intensity distribution in the interference field is determined both by the amplitudes and phases of interfering waves.

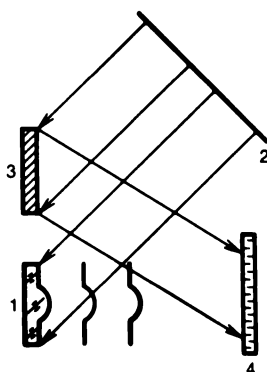
Consequently, interference of light can be used for recording all characteristics of the wave field propagating from an object of observation if the conditions required for interference are ensured.

A time-independent interference pattern is formed as a result of interference of coherent light waves. Thus, for recording phase distribution in the wave field obtained in the presence of an object of observation, one should first of all illuminate the object by a monochromatic and spatially coherent radiation. Then the field scattered by the object will possess these properties as well.

If we now superimpose on the field produced by the object an auxiliary monochromatic field of the same frequency, say, a plane wave (the so-called *reference wave*), a complex, time-invariant distribution of the regions of mutual amplification and attenuation of the two waves, i.e. a stationary interference pattern, is formed in the entire region of space where the two waves (the one scattered by the object and the reference wave) overlap. Such a constant intensity distribution of the resultant field can then be recorded on a photographic plate. Naturally, the intensities at the points of space lying in the plane of the plate will be recorded on the plate.

---

<sup>1</sup> Parallactic displacement is the visible shift in the mutual arrangement of objects of observation as a result of a change in the position of an observer.



**Fig. 283.**  
Schematic diagram of a holographic  
recording of an opaque object.

The schematic diagram of a holographic set-up for the recording of an opaque object of observation *1* in reflected light is shown in Fig. 283.

Here *2* is the front of a plane wave produced by a laser beam which is diverged to the required cross section by a special optical system.

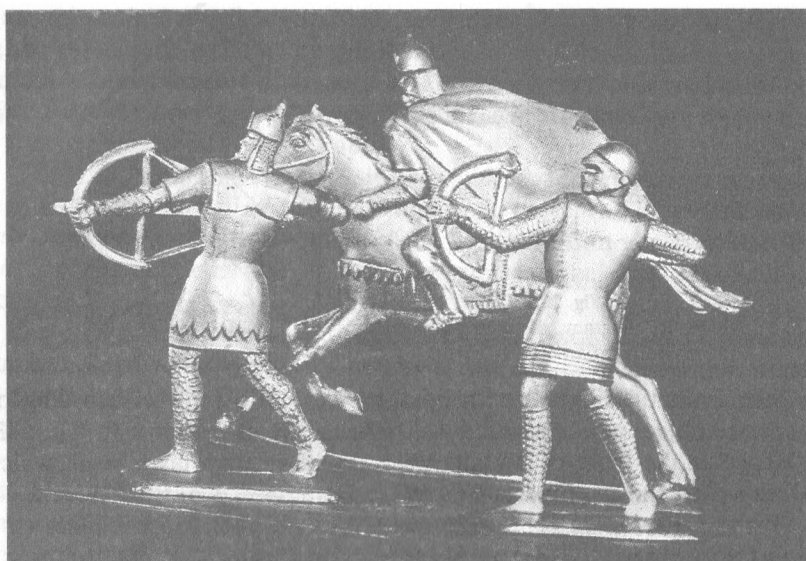
Mirror *3* directs the plane reference wave to a photographic plate *4* on which the waves scattered by object *1* are incident.

It is essential that along with the reference wave, the light waves scattered by all regions of the object of observation are incident on each point of the illuminated part of the photographic plate. For this reason, any element of a hologram contains the complete information about the entire object of observation.

The recording of holograms in conformity with the above scheme imposes certain requirements on the spectral composition of the radiation used in this case. Indeed, for the emergence of interference patterns as a result of superposition of the wave field propagating from the object of observation and the field of the reference wave, these fields must be coherent for any propagation difference. The propagation difference is inevitable because of the macroscopic relief of the object reflecting light, and they may be significant.

If, for example, we assume that the propagation difference attains 10 cm, calculations show that the spectral width of the employed radiation must be of the order of  $10^{-2}$  Å. Meanwhile, the spectral width of the light emitted by a mercury-vapour lamp is of the order of tens of angstroms even for a high vapour pressure. Consequently, the light sources with the so-called "prelaser" period of evolution of optics are almost useless for holography. On the other hand, using lasers it is not difficult to satisfy the requirements to light monochromaticity set by holography. This explains a rapid development of holography in recent time when lasers have become available in many laboratories.

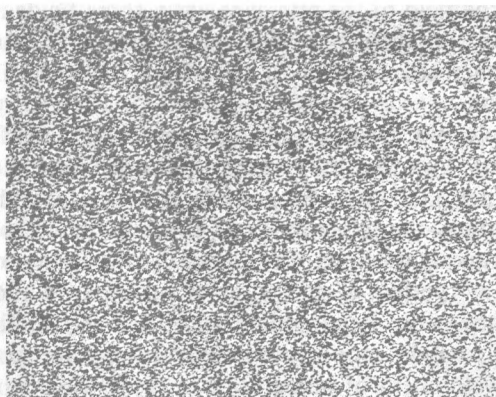
Another important aspect is noteworthy. In recording holograms of ex-



**Fig. 284.**

The image of an object was obtained by ordinary photography.

tended objects, the angles between interfering light waves incident on a photographic plate may reach significant values. Therefore, the interference pattern formed on the photographic plate turns out to be very small, and photographic materials with a large resolution are required for its recording. Modern holographic plates have a resolving power of more than 5000 lines per millimetre.



**Fig. 285.**

Magnified image of a region of a hologram.

The following two photographs can illustrate the first stage of holographing. The first of them (Fig. 284) shows the image of objects, obtained by taking an ordinary photograph, while the second (Fig. 285) shows the photographic recording of the interference pattern under a large magnification, viz. a hologram of the objects, recorded with the help of a plane reference wave. It can be seen that the photographs are quite different.

### 15.3. Obtaining Optical Images by Reconstructing the Wave Front

In order to decode the information about an object of observation recorded on a hologram obtained by the method described above, the hologram is illuminated.

The schematic diagram of a set-up intended for obtaining holographic images is shown in Fig. 286.

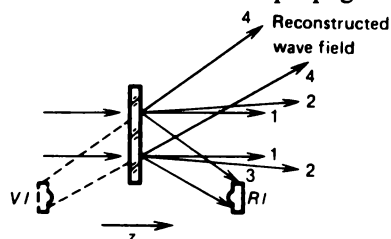
A plane monochromatic wave is incident on a hologram from the left normally to its surface. As a result of the passage of light through the hologram and its diffraction on nonuniform blackening, a complex system of light beams is formed behind the plate.

Beams 1 and 2 carry no information about the object. A converging light beam 3 forms a real image *RI* of the object without any optical system. It is formed at the same distance from the hologram at which the object of observation was relative to the photographic plate during the recording.

By arranging a screen at different cross sections of the region of localization of the real image, one can observe on the screen sharp images of various details of the object. For fixing them, it is sufficient to use a photographic plate instead of a screen. When different parts of a hologram are illuminated, the effect of mutual parallax displacement of the details of the object is observed since the light from the object was incident on these details at different angles during holographic recording.

A diverging light beam 4 passed through a converging lens reconstructs another image of the object, viz. the so-called virtual image *VI*. It is localized in front of the hologram symmetrically to the real image. The virtual image can also be observed by unaided eye. The role of converging lens is played in this case by the eye lens which projects the image on the retina.

Thus, in the absence of the object, the same wave field which propagat-



**Fig. 286.**

Schematic diagram of the reconstruction of holographic images.

ed from the object during the holographic recording is reconstructed behind the illuminated hologram. As a result, the object can be photographed or examined from different points in space as if the object were located in front of the observer or camera.

As was mentioned above, the wave field reconstructed by a hologram makes it possible to observe and register the effects of parallaxic displacements of the details of the object. For this purpose, one should either change the mutual spatial orientation of the hologram and the light beam illuminating it, or change the position of the eye (or the objective of the camera) relative to the stationary hologram as is done in walking around the object or a group of illuminated objects for viewing them from different sides.

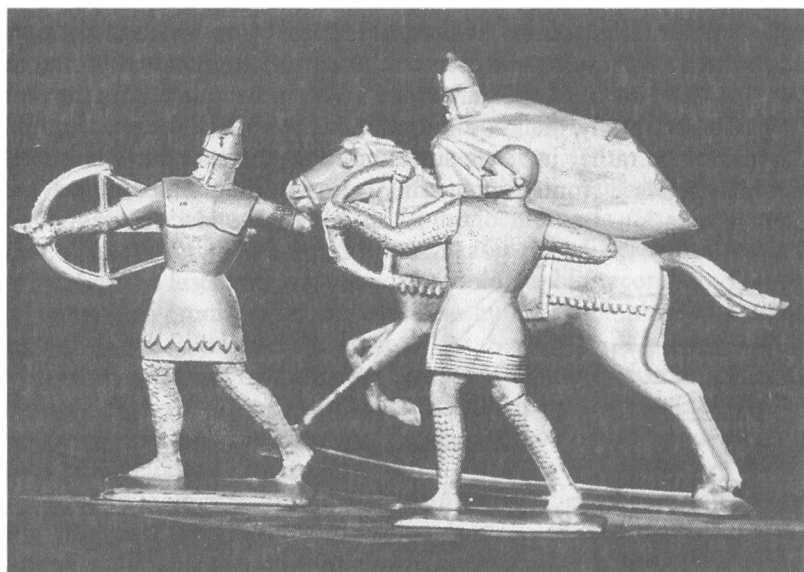
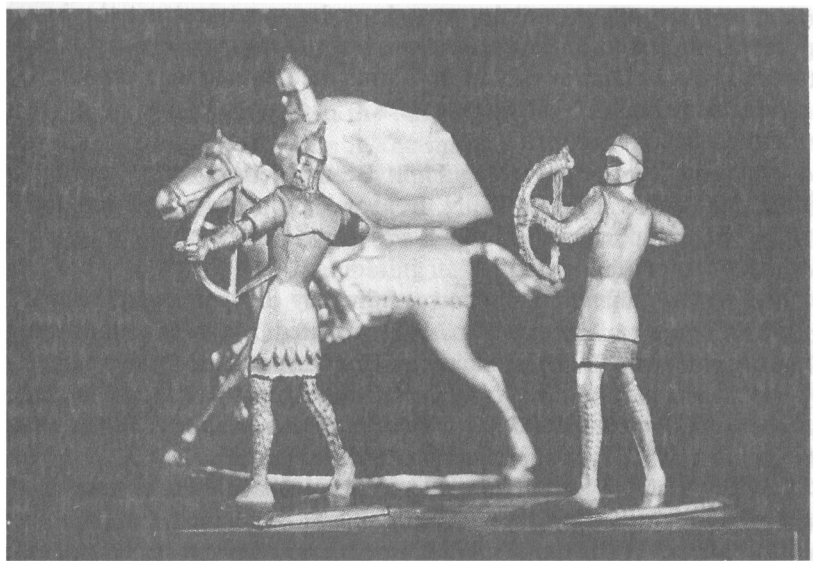
The effect of parallaxic displacement is illustrated by Fig. 287. The arrangement of objects on the photographs is as if viewed from different points. Meanwhile, the two images were obtained with the help of the same hologram, but the camera was placed in different positions relative to it. Thus, as was mentioned above, a hologram contains much more information about an object than a conventional photograph.

We shall indicate some more important properties of the holographic method of recording and reconstruction of optical information.

In order to obtain optical images by illuminating a hologram, we can use only a part of its area. By illuminating any part of the hologram, we can completely reconstruct the real and virtual images of the object. This is a direct consequence of the fact that light waves scattered by all elements of the surface of the object, as well as the front of the reference wave, reach any region of the surface of the photographic plate on which the hologram is recorded. We must only take into account the fact that a considerable reduction of the area of a hologram used for reconstructing leads to a reduction of its resolution, and the image of the object is blurred, i.e. it becomes difficult to distinguish small details of the object.

Besides, the concepts of the positive and the negative do not exist in holography. Using a replica prepared from a hologram by the contact method, it is possible to reconstruct images with the same distribution of light and shadow regions as in the primary hologram. Naturally, this is strictly observed only for two-dimensional diffraction patterns. In actual practice, the thickness of the photographic emulsion layer is about  $20\text{ }\mu\text{m}$ , i.e. about 100 wavelengths of light fit into it. For this reason, some effects characteristic of a three-dimensional structure are manifested even when photographic plates with thin emulsion layers are used. To a certain extent, this circumstance sets a limit to the hologram duplicating by contact printing.

Finally, the same plate can be used for recording holograms of several



**Fig. 287.**  
The parallax displacement observed for holographic images.

objects by varying the plate orientation relative to the wave fields being recorded and the reference light wave. In order to reconstruct the images of different objects recorded on the same hologram separately and without distortions, the hologram should be illuminated by monochromatic light beams incident on it at different angles.

#### 15.4. Holographing by Opposing Light Beam Method

In 1962, the Soviet physicist Yu. N. Denisjuk proposed a new method of obtaining holographic images, which was based on the obsolete Lippman coloured photography method.

The diagram of a holographic recording by this method is shown in Fig. 288. An object of observation 1 is illuminated by a laser beam through a photographic plate (holographic plates are transparent to light even before they are developed and fixed). A glass film support of plate 2 is coated by a photographic emulsion layer with a thickness of about  $15\text{--}20\text{ }\mu\text{m}$  (this layer has a much larger thickness in the figure).

The wave field reflected from the object propagates backward in the direction to the emulsion layer. The initial light beam 4 from the laser propagates from the opposite direction and now serves as the reference wave. For this reason, this version of obtaining holograms is also known as the *opposing light beam method*.

The interference field of standing waves emerging in the bulk of the photographic emulsion causes its laminated blackening which records amplitude as well as phase distribution of the wave field scattered by the object. In Fig. 288, the blackened layers are shown schematically in the form of wavy lines. Naturally, their configuration in the bulk of the photographic emulsion may be rather intricate since only the reference wave is a plane wave while the wave fronts propagating from the illuminated object have random orientation.

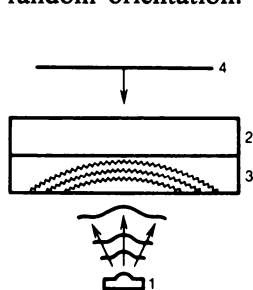


Fig. 288.

Schematic diagram of a holographic recording by the opposing light beam method.

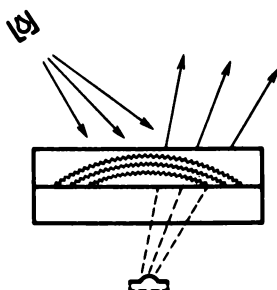


Fig. 289.

Schematic diagram of the reconstruction of a virtual image.

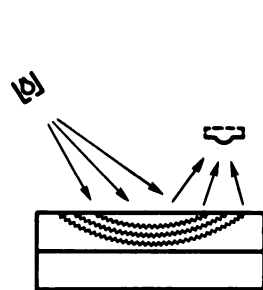


Fig. 290.

Schematic diagram of the reconstruction of a real image.

It is essential that the thick-layered emulsion with a nonuniform blackening distribution is a *three-dimensional* structure unlike two-dimensional structures of the holograms of the type considered earlier (to a high degree of approximation).

If a hologram recorded in the thick-layered emulsion is illuminated by a diverging beam of white light, the image of the object can be observed in the light reflected from the hologram.

Figures 289 and 290 illustrate how a real or virtual image of an object of observation can be obtained by varying the orientation of the hologram relative to the light beam illuminating it. Naturally, the reconstructed image will show not the entire opaque object but only its surface facing the photographic plate during the hologram recording.

It is possible here to use sources of white light (like the Sun or incandescent lamps) at the stage of reconstruction of the image recorded in a thick-emulsion hologram due to the fact that the mutual amplification of light beams reflected from the regions of blackening in the three-dimensional diffraction structure having a definite spatial period at a certain angle of vision is observed only for a radiation at a certain wavelength. Thus, spatial periodic layers of three-dimensional diffraction structure automatically ensure the radiation monochromatically required for observing a holographic image. The image is reconstructed in a monochromatic light.

The spectral resolution of a three-dimensional diffraction grating with a small number of blackened layers is obviously insufficient for the monochromatization of the white light illuminating a hologram to the same extent as the monochromaticity of laser radiation used in the stage of the hologram recording. For this reason, images produced by thick-emulsion holograms are not completely monochromatic.

Besides, although the images obtained by illuminating thick-emulsion holograms by white light are quasi-monochromatic (i.e. not completely monochromatic), their colours in some cases may differ considerably from the colour of laser radiation used for holographic recording. This is due to the effects produced on the emulsion by the developing, and especially by fixing and subsequent drying.

Another feature of holograms recorded on thick-layer emulsions with the help of the opposing beam method is worth mentioning. It is associated with the effect of *pseudoscopy* characteristic of holography, which is exhibited in this case most clearly.

If a hologram is recorded according to the scheme represented in Fig. 288, and the image of the object is reconstructed by illuminating the hologram in accordance with the diagram shown in Fig. 289, the virtual image of a convex object will also be convex. On the other hand, in the real image of the object (Fig. 290), the convex surface of the object will



be concave since the parts of the object lying closer to the photographic plate during hologram recording will now lie closer to the hologram.

For this reason, virtual images are used for observing museum exhibits which are seen behind the hologram plane.

Hologram recording in thick-layer emulsions makes it possible to obtain coloured images of objects, other advantages of holography over photography being preserved.

In order to explain the principle of coloured holography, it should be recalled in what cases the eye perceives images of object as coloured, and not as black and white.

Experiments in vision physiology indicate that an image is perceived by the human eye as coloured and more or less close to natural colours if it is reconstructed at least in three colours, say, red, green and indigo. The combination of the three colours is used in the most primitive coloured reproduction of a picture obtained by lithographic methods (for high-quality reproductions, 10-15-coloured printing is used).

Taking into account the peculiarities of human perception, one can reconstruct a coloured image of an object by illuminating it during a hologram-recording simultaneously or consecutively by laser radiation with three spectral lines separated by sufficiently large wavelength regions. Then three systems of standing waves are formed in the bulk of the emulsion, and accordingly there are three systems of spatial lattices with different blackened distributions.

Each system of darkening layers will form an image of the object in its own spectral region of white light, used in reconstructing the image. Owing to this, the image of the object will be formed in the diverging beam of white light reflected from the processed hologram, as a result of superposition of the three spectral regions. This corresponds to minimum physiological requirements of chromatic vision.

The Denisjuk method of holography and image reconstruction according to the diagram shown in Fig. 289 are widely used for obtaining high-quality three-dimensional copies of various objects like unique pieces of art.

### **15.5. Application of Holography to Optical Interferometry**

It is well known that interference of light has found many diversified applications in physics and engineering. For example, interference of light is widely used in high-precision control of the geometrical shape of various bodies, the quality of finishing their surfaces, small variations of shape or surface under the influence of various external effects like mechanical stresses, or heating.

However, in ordinary optical interferometry an object under investiga-

tion is compared with a specially prepared standard, and the surface of the object as well as that of the standard must be thoroughly finished.

Holography can be used for the interferometric analysis of, say, deformations of an arbitrary-shaped body with any surface finishing.

Let us consider again Fig. 283 representing the diagram of holographing objects of any shape. Let us suppose that we have to investigate, by using the methods of holographic interferometry, small deformations of an object caused by some reason or other.

We expose a holographic plate, illuminating the object prior to the deformation. Without displacing the plate and developing it, we interrupt the illumination of the object for an arbitrary time interval. During this intermission, we deform the object without changing its position in the holographic set-up. Then we illuminate the deformed object again and expose the holographic plate once more. After the second exposition, the holographic plate is developed and fixed as usual.

As a result, we shall obtain on the plate two holograms recorded with the same reference wave. The first will be the hologram of the undeformed object, and the second will be that of the deformed object. As was mentioned earlier, it is quite admissible to record two or more holograms on the same plate (unlike the case of several overlapping optical images on an ordinary photographic plate).

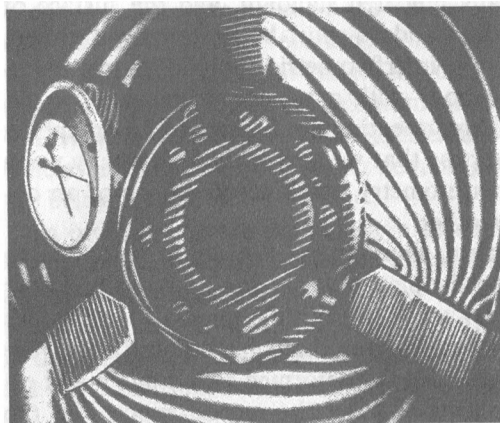
The images of the deformed and undeformed object with the help of a "double" hologram are reconstructed in accord with the diagram shown in Fig. 286. As was explained above, only light beams 3 and 4 which will be considered carry the complete amplitude and phase information.

Since in this case a "double" hologram is illuminated, two wave fields 3 and 3', and 4 and 4' will emerge behind it. Each pair of these fields corresponds to the deformed and undeformed object respectively.

Since the two pairs of fields we are interested in are formed as a result of illuminating the hologram with the same spatially coherent light beam, the wave fields of each pair may interfere with each other to form a stable interference pattern. But the difference between the wave fields 3 and 3' (as well as between 4 and 4') is that the object has been deformed in the interval between the two holographic recordings. Consequently, as a result of illumination of the "double" hologram, the superposition of fields 3 and 3', as well as 4 and 4', will produce in the real and virtual images interference patterns revealing the nature of deformation.

It is essential that in this method the standard for comparing with the deformed object is the object itself rather than a specially prepared surface finished to within the optical accuracy.

Thus, holographic interferometry drastically broadens the possibilities of optical interferometric observations and measurements. In particular,



**Fig. 291.**  
The image of a deformed ball bearing.

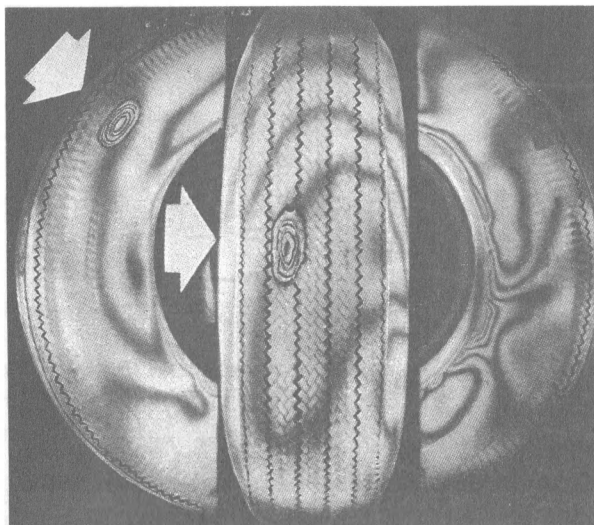
as was shown above, it makes it possible to investigate deformations of objects with arbitrary shape and quality of the surface as well as details which are not prepared especially in any way.

Figure 291 shows an image of a ball bearing clenched in the jaws of a lathe chuck. It was obtained as a result of illuminating the “double” hologram recorded first before and then after the emergence of deformations in the object. The interference fringes on the surface of the ball bearing reveal the strain distribution.

Holographic interferometry is widely used for indestructible control. For example, it is possible to reveal cavities and weak points in welded parts of the walls of hollow vessels. For this purpose, air inside a vessel is heated, which causes the expansion of the walls. The regions having different thermal conductivities expand differently. The pattern of interference fringes allows one to reveal regions in which the thermal conductivity differs from the normal value. Similarly, vessels under pressure can be tested in the same way: the weakened regions will be covered by denser interference fringes.

Another application of holographic interferometry is the control of the quality of motor-car tyres from the deformation of their surfaces as a result of a slight change in pressure (Fig. 292). Weakened regions are characterized by a high concentration of interference fringes (shown by the arrows). The figure represents reconstructed images in two projections.

Holographic interferometry can also be used for studying vibrational processes. In this case, a hologram of a vibrating surface (say, of a membrane) is recorded in a conventional way, so that the exposure time is considerably longer than the period of vibrations of the membrane. Thus, during the exposure time the vibrating surface passes many times through all the positions between the two limiting ones. However, most of the time

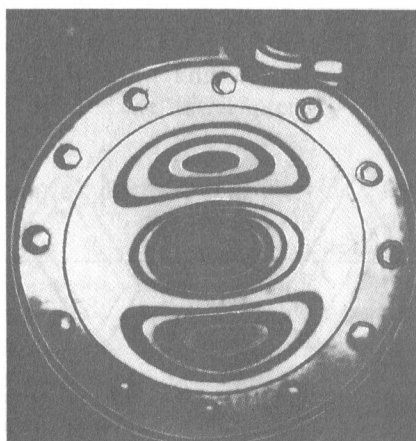


**Fig. 292.**

The image of a damaged motor-car tyre (in two projections).

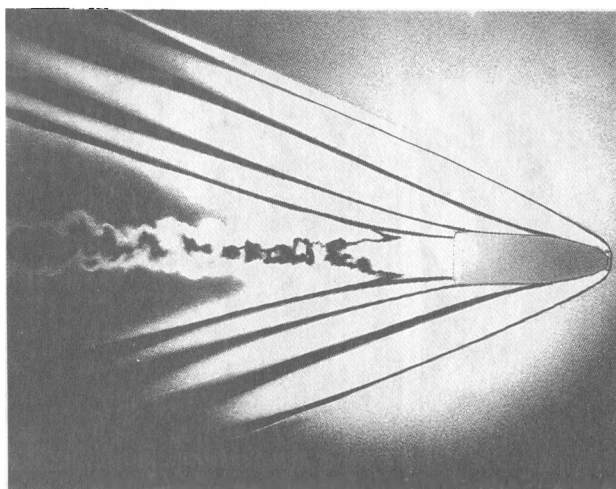
the membrane is in the limiting positions since the velocity of motion of the membrane is minimum when the deviation from the equilibrium position is maximum.

The obtained time-averaged hologram can be treated as a double exposure hologram. As in the case of the two-exposure method, it can be



**Fig. 293.**

The image of a vibrating membrane.



**Fig. 294.**

The image of shock waves produced by a flying bullet.

used for observing interference fringes allowing one to calculate the amplitude of vibrations for different points of the membrane. The image of a membrane reconstructed from such a hologram is shown in Fig. 293.

The double-exposure method using lasers which produce high-power short light pulses can be successfully used in interferometry of rapid processes. Figure 294 represents the image of a flying bullet obtained with the help of a double-exposure hologram. One can see interference fringes in the region of the shock wave.

## Chapter 16

# Polarization of Light. Transverse Nature of Light Waves

### 16.1. Passage of Light Through Tourmaline

Interference and diffraction, which substantiated the wave nature of light, still do not give a full idea about the nature of light waves. New properties of light are illustrated by the experiment with light passing through crystals like tourmaline.

Let us take two identical rectangular tourmaline<sup>1</sup> plates cut so that one of the sides of the rectangle coincides with a certain direction in the crystal, which is known as the *optical axis*.

We place one plate on the other so that their axes coincide in direction and pass a narrow light beam from a lantern or the Sun through this pair of plates. Since tourmaline is a greenish-brown crystal, the trace of the passing beam on a screen has the form of a dark-green spot. Let us rotate one of the plates about the beam, leaving the other plate at rest (Fig. 295). We shall see that the trace becomes weaker and weaker, and vanishes when the plate turns by  $90^\circ$ . A further rotation of the plate leads to an increase in the light intensity which attains the initial value when the plate is rotated through  $180^\circ$ , i.e. when the optical axes of the plates become parallel again. When the tourmaline crystal is rotated further, the beam attenuates again, passes through the minimum (vanishes) when the axes of the plates are perpendicular, and reaches the initial intensity when the plate returns to the initial position.

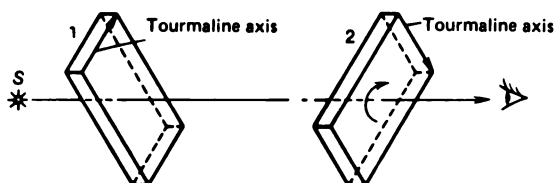


Fig. 295.

Schematic diagram of the experiment on the passage of light through two tourmaline plates:  
S — light source, 1 and 2 — the first and the second tourmaline plates.

<sup>1</sup> Tourmaline is a single crystal with a complex chemical composition (containing oxides of aluminum, silicon, boron and other elements). — Eds.

Thus, when the plate is turned through  $360^\circ$ , the intensity of the beam passing through the two plates attains the maximum value twice (when the axes of the plates are parallel). The process occurs in the same way irrespective of which of the two plates is rotated and of the direction of rotation. It is no matter whether the plates are in contact or at a certain distance from each other (Fig. 295).

If, however, we remove one of the plates and rotate the remaining one, or rotate the two plates so that their axes remain at a constant angle, no change in the passing beam intensity will be observed. Thus, the intensity of the passing beam changes only when the light passing through one of the plates encounters the other plate whose axis changes its direction relative to the axis of the first plate.

## 16.2. Hypotheses Explaining Observed Phenomena.

### Polarized Light

Thus, light passing through tourmaline acquires special properties. The properties of light waves in the plane perpendicular to the direction of propagation become *anisotropic*, i.e. different relative to the plane passing through the light beam and the tourmaline axis. For this reason, the ability of such a light to pass through the second tourmaline plate depends on the orientation of the optical axis of this plate relative to the optical axis of the second plate. The light beam emitted by the lantern (or the Sun) did not exhibit such an anisotropy since the tourmaline orientation was immaterial for this beam.

The observed phenomena can be explained if we draw the following conclusions.

1. Light oscillations in the beam are directed at right angles to the direction of light propagation (light waves are transverse).
2. Tourmaline can transmit light oscillations only if they are oriented in a certain way relative to its axis (say, parallel to it).
3. In the light emitted by the lantern (the Sun), oscillations of any direction are represented equally so that there is no preferential direction.

Henceforth, the light in which all directions of transverse oscillations are represented equally will be called a *natural light*.

Conclusion 3 explains why natural light passes through tourmaline equally well for any orientation of the crystal, although, according to conclusion 2, tourmaline transmits light oscillations only of a definite direction. Indeed, irrespective of the tourmaline orientation, the natural light always contains the same fraction of oscillations whose direction coincides with that transmitted by the tourmaline crystal. The passage of natural light through tourmaline results in the selection of only those transverse oscillations which are transmitted by the tourmaline crystal. For this reason, the

light that has passed through the tourmaline crystal is an aggregate of oscillations having the same direction determined by the orientation of the tourmaline crystal axis. Such a light will be termed *linearly polarized*, while the plane containing the direction of oscillations and the axis of the light beam is known as a *polarization plane*.

Now we can explain the experiment with the passage of light through two consecutive tourmaline plates. The first plate polarizes the light beam passing through it, leaving in it oscillations of only one direction. These oscillations can completely pass through the second tourmaline plate only if their direction coincides with the direction of oscillations transmitted through the second plate, i.e. when its axis is parallel to that of the first plate. If the direction of oscillations in a polarized light is perpendicular to the direction of oscillations transmitted by the second tourmaline plate, the light will not be transmitted at all. This is the case when the plates are crossed, i.e. their axes are at  $90^\circ$ . Finally, if the direction of oscillations in polarized light forms an acute angle with the direction transmitted by the tourmaline plate, the oscillations will be transmitted only partially.

### 16.3. Mechanical Model of Polarization

The explanation provided in the previous section can be illustrated with the help of mechanical experiments. A rope vibrating in one plane, say, vertical, may serve as a model of a polarized light wave. A model of natural light wave is a rope whose vibration plane rapidly changes, assuming different orientations in a short time. Two boards separated by a narrow gap (slit) play the role of a model of tourmaline plate: the rope vibrations directed along the slit easily pass through the gap, while the vibrations perpendicular to the slit are not transmitted. The experiments represented in Fig. 296 completely correspond to the optical experiments described above. They show that “natural” vibrations of the rope are equally transmitted for any orientation of the slit. Two consecutive slits transmit vibrations of larger or smaller amplitude depending on their mutual orientation. When the slits are at right angles, they do not transmit vibrations of the rope. Experiments also show that a slit polarizes “natural” vibrations of the rope.

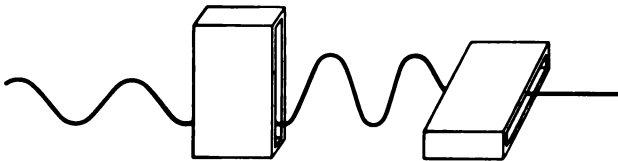


Fig. 296.

Mechanical model of the passage of a light wave through two tourmaline plates.



### 16.4. Polaroids

A tourmaline crystal is far from being a single crystal polarizing light passing through it. A large number of crystals exhibit similar properties. However, most of them (like Iceland spar) transmit simultaneously two rays which are polarized in two mutually perpendicular directions. This often deteriorates observations of polarized light and necessitates special appliances for separating these rays. Some crystals, including tourmaline, absorb one of the two polarized rays so strongly that practically only one ray passes through a plate about one millimetre thick, which is polarized in a definite direction. Such crystals are known as *dichroic*.

There are crystals that absorb one of the polarized rays even stronger than tourmaline (for example, quinine iodide) so that a crystalline film whose thickness is a tenth of a millimetre or even smaller separates one of the polarized rays almost completely. Coating celluloid by such films, we can obtain a polarizing plate of a few square decimetres. Such plates are called *polaroids* and are convenient and cheap polarizing appliances of a large surface. All experiments described in Sec. 16.1 can easily be carried out with two pieces of a polaroid.

### 16.5. Transverse Nature of Light Waves and Electromagnetic Theory of Light

The hypotheses contained in Sec. 16.2 provided such a complete and good explanation to all details of the experiments with tourmaline that they can be regarded as well substantiated. The most important of them concerns the transverse nature of light waves. Using the idea of transverse light waves, it is also possible to successfully explain many other phenomena associated with polarization of light. Thus, a large and diversified group of polarization phenomena serves as a reliable substantiation of the idea according to which *any light wave is a transverse wave*, i.e. directions of oscillations in it are perpendicular to the direction of its propagation.

The fact that light waves were accepted to be transverse was of major importance in the theory of light. Fresnel, Young<sup>2</sup> and other scientists who substantiated the wave nature of light assumed that light waves are similar to elastic waves propagating in a special sort of medium filling the entire space and called the light ether. Later, however, it turned out that the hypothesis of elastic ether and the concept of light as elastic waves could not provide a satisfactory explanation to a number of newly discovered phenomena. For example, some facts were established which revealed a close relationship between electromagnetic and optical phenomena. Among them, in the first place there were experiments demonstrating the possibility

---

<sup>2</sup> Thomas Young (1773-1829) was the British physicist and physician.

to affect the nature of light polarization with the help of magnetic or electric fields. Further the effect of electric and magnetic fields on the frequency of light emitted by atoms was discovered, as well as the possibility to cause some electric processes with the help of light (e.g., photoelectric effect, see Sec. 21.2), and so on. The relation between optical and electromagnetic phenomena was reflected in the electromagnetic theory of light proposed by Maxwell in 1876 (see Sec. 6.5).

The electromagnetic theory of light has removed all difficulties introduced by the hypothesis of elastic ether. In order to explain the process of propagation of electromagnetic waves, there is no need to assume that the world space is filled with any substance. Electromagnetic waves (including light) can also propagate in vacuum (cf. Sec. 4.1). An electromagnetic wave is (see Secs. 6.1 and 6.6) the propagation of varying electromagnetic field in which the electric field strength and magnetic induction vectors are perpendicular to each other and to the direction of wave propagation: *electromagnetic waves are transverse*. Thus, the transverse nature of light waves proved in experiments on polarization of light is explained in a natural way by the electromagnetic theory of light. As in any electromagnetic wave, in a light wave there are simultaneously two mutually perpendicular directions of oscillations, viz. the directions of oscillations of electric and magnetic fields. All that was said about the direction of light oscillations refers to the direction of oscillations of electric field strength. In particular, special experiments made it possible to establish that in a wave passing through a tourmaline crystal the oscillations of the electric field strength are directed along the optical axis of the crystal.

## Chapter 17

# Electromagnetic Spectrum

### 17.1. Methods of Investigating Electromagnetic Waves of Different Wavelengths

Electromagnetic waves used in radio engineering have wavelengths from several kilometres to a few centimetres. On the other hand, electromagnetic waves representing light are characterized by wavelengths of the order of tenth of a micrometre. This simple comparison shows that the quantitative difference in wavelength results in a deep qualitative difference in many properties and features of electromagnetic waves. It is very important to analyze the properties of electromagnetic waves of different wavelength in greater detail. In order to separate waves with different wavelengths, one of the methods of spectral decomposition of a complex radiation is usually employed. For visible light, a diffraction grating (see Sec. 14.9) or a prism (see Sec. 9.8) can be used for this purpose.

Analyzing a spectrum obtained on a screen, we see that it is possible to distinguish waves having different wavelength by eye. However, as was mentioned more than once, the eye perceives only electromagnetic waves having wavelengths approximately between 400 and 760 nm. Naturally, these boundaries are rather indefinite and some observers can “see” waves slightly shorter (to about 370 nm) or longer (about 800 nm). Therefore, it is necessary to find a more general method for detecting electromagnetic waves than observation by eye.

Since a propagating electromagnetic wave of any wavelength carries an energy, such a more general method can be associated with measuring the energy of the wave. The most convenient technique consists in conversion of the electromagnetic energy of the wave into the internal energy of the substance; an increase in the value of this energy is accompanied by heating of the body. Heating of bodies can easily be observed with the help of sensitive thermometers like thermocouples (see Vol. 2, Sec. 6.10). A partial conversion of energy of electromagnetic waves into the internal energy takes place every time when these waves are incident on a substance and are absorbed in it to a certain extent. Experiments revealed that some black materials like carbon black absorb the energy carried by light waves of different wavelengths almost completely. For this reason, they appear as black, i.e. not reflecting light.

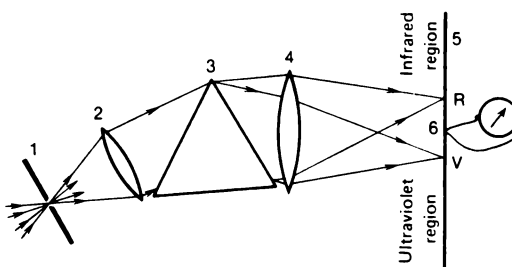


Fig. 297.

Schematic diagram of the experiment on the energy distribution in a spectrum: 1, 2, 3, 4 — parts of a spectrometer producing the spectrum of a source in plane 5; 6 — thermocouple that can be moved along the spectrum, 7 — galvanometer, V — the violet boundary of the spectrum, R — the red boundary of the spectrum.

Having covered by carbon black the sensitive part of a thermocouple and moving it over the spectrum we can investigate electromagnetic waves in a wide wavelength range. Figure 297 shows the arrangement of the elements of optical system intended for this purpose. Having measured the degree of heating of the thermocouple, we can calculate the energy corresponding to a certain spectral region, i.e. judge about the energy distribution in the spectrum. Such measurements of energy give results differing from conclusions drawn by the eye. Indeed, for a person perceiving light by the eye it seems that the yellow or green part of the spectrum of an arc lamp is much brighter than the red one, while a thermocouple indicates stronger heating in the red region. The reason behind this lies in the peculiarities of the eye whose sensitivity to different colours is different (see Sec. 8.1). For this reason, the eye fails to give a right idea about the energy distribution in a spectrum. On the other hand, the thermocouple is quite an “impartial” instrument since it allows us to judge about the internal energy into which the light energy is converted as a result of absorption for any wavelength.

### 17.2. Infrared and Ultraviolet Radiation

An analysis of the energy distribution in a spectrum reveals that readings of the thermocouple do not vanish when it is moved to the region where the eye does not see anything, i.e. when the instrument is placed behind the red or violet boundaries of the spectrum (see Fig. 297). The readings of the thermocouple gradually change as we go over to these invisible regions of the spectrum. For many sources (like an arc lamp), the readings of the thermocouple even become higher as we move to the region behind the red boundary of the spectrum in spite of the fact that the eye sees no light in this region. As we go over to the region corresponding to still longer waves, the readings of the thermocouple become smaller. The waves having

longer wavelength than red waves are known as *infrared*. They were discovered by the English physicist and astronomer John Herschel (1792-1871) in 1830 during the analysis of energy distribution in a spectrum with the help of a very sensitive thermometer. The waves whose wavelengths are larger than those of violet waves are known as *ultraviolet*.<sup>1</sup> Since the energy corresponding to the violet and ultraviolet spectral regions is not high for conventional sources, it is rather difficult to investigate this spectral region with the help of a thermocouple, although this technique is always used for accurate measurements of energy.

It is much easier to detect ultraviolet waves from their effect on a photographic plate or paper. Having directed the light from a lantern, decomposed into spectral components, onto a strip of photographic paper,<sup>2</sup> we shall see that the paper soon becomes black in the regions of blue and especially violet waves, while under the action of the green and yellow parts of the spectrum it remains white. A still stronger blackening is observed behind the violet region. In 1801, the English physicist William Wollaston (1766-1828) observed ultraviolet waves in similar experiments with  $\text{AgCl}$ .<sup>3</sup> A convenient method of detecting ultraviolet radiation is based on fluorescence and phosphorescence (see Sec. 21.8).

### 17.3. Discovery of X-rays

X-rays were discovered in 1895. The method of obtaining them visually demonstrates their electromagnetic nature.

The German physicist Wilhelm Roentgen (1845-1923) discovered this type of radiation accidentally while investigating cathode rays (see Vol. 2, Sec. 8.12). The diagram of the experiment was as follows. The tube for obtaining cathode rays had the form shown in Fig. 298. When a discharge was passed through the tube, he observed the following phenomenon. A piece of paper coated with barium platinum cyanide<sup>4</sup> flashed brightly (fluoresced) during each discharge as it was brought close to the tube enclosed in black cardboard. It could be concluded that black cardboard impenetrable to visible or ultraviolet solar rays is pierced by an agent causing its intense phosphorescence. In a series of experiments, Roentgen estab-

---

<sup>1</sup> Infrared radiation is that whose frequency is lower than the frequency of the visible light. The frequency of ultraviolet radiation is higher than that of visible light.

<sup>2</sup> Naturally, the so-called *daylight photographic* paper should be used in experiment, which gets dark in light without any treatment (developing).

<sup>3</sup> Naturally, in these experiments a silver chloride solution blackened under the action of light was used instead of photographic paper which was not known at that time.

<sup>4</sup> A barium platinum cyanide layer glows if it is preliminarily illuminated by visible or ultraviolet light. This glowing is known as *phosphorescence*.

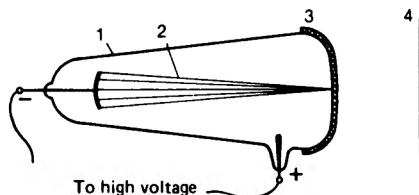


Fig. 298.

To the discovery of X-rays. Gas-discharge tube 1 for experiments with cathode rays 2 is covered by a black cardboard casing 3. The glow is observed on phosphorescent screen 4.

lished that this “agent”, called by him X-rays<sup>5</sup>, also passes through other bodies that are opaque for ordinary light: paper, wood, ebonite, human body and metal layers. Roentgen also found out that low-density materials consisting of light atoms are more transparent for X-rays than materials with a higher density. For example, a lead plate absorbs X-rays much stronger than an aluminium plate of the same thickness. The bones of a human body absorb these rays stronger than flesh. Therefore, having placed a hand between the source of X-rays and the screen, we shall see a weak shadow of the hand in which dark shadows cast by bones can easily be seen (Fig. 299).

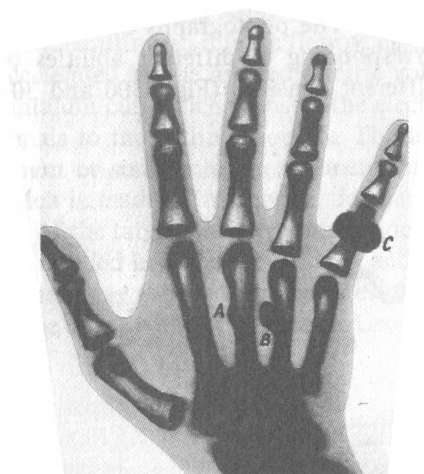
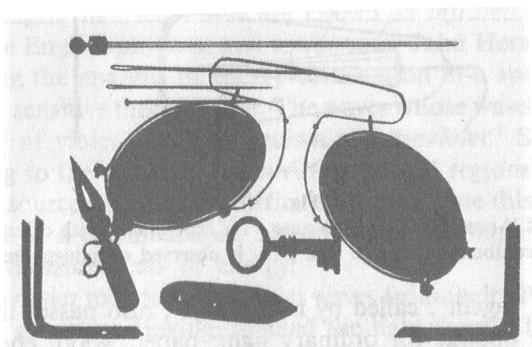


Fig. 299.

X-ray photograph of the palm of a hand: A, B — fragments of a bullet, C — ring on the little finger.

<sup>5</sup> The term “X-rays”, i.e. unknown rays, was used by Roentgen till his last days. Other scientists named them Roentgen rays.

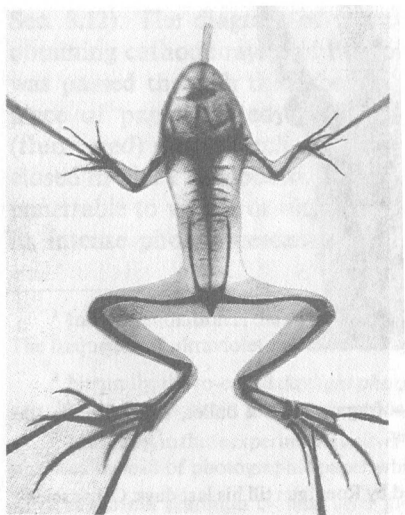


**Fig. 300.**

X-ray photograph of a purse containing some metal objects: the glasses of the pince-nez are made of glass containing lead. For this reason, they do not transmit X-rays.

#### **17.4. Effects of X-rays**

After the first experiments in which the ability of X-rays to cause phosphorescence was revealed, their other properties were also discovered. X-rays may cause chemical processes. For example, acting on a photographic plate or paper, they cause blackening. This phenomenon forms the basis of the X-ray photography. The photographs obtained are of the shadow type with details corresponding to different abilities of X-rays to pass through bodies of different densities (Figs. 300 and 301).



**Fig. 301**

X-ray photograph of a frog: the bones of the skeleton can clearly be seen. The legs are pinned to the support by metal pins.

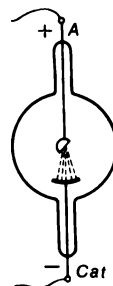
These properties of X-rays have found important applications in medicine and engineering. Using X-rays, it is possible to observe defects or changes in the bulk of an object on a phosphorescent screen or photographic plate (like the defects in the parts of machines, changes in human organism, etc.). The property of X-rays to cause chemical changes can be used for curing organs damaged by some types of disease (like cancer). Here it is important that it is possible to influence by X-rays the functions of internal organs of a living organism.

It is interesting to note that some grades of glass containing lead compounds are penetrable to visible rays but strongly absorb X-rays (Fig. 300), while ordinary glass (containing sodium salts) is transparent both for visible light and X-rays.

### 17.5. X-ray Tube

Having discovered X-rays, Roentgen carried out thorough experiments to clarify the conditions for their formation. He established that these rays emerge in the region of the tube where flying electrons constituting cathode rays are stopped as a result of collision with the tube wall. Proceeding from this fact Roentgen constructed a special tube convenient for obtaining X-rays. The main features of this construction can be found in modern X-ray tubes.

Figure 302 represents a modern X-ray tube. The cathode is a thick red-hot tungsten filament emitting an intense flow of electrons (see Vol. 2, Sec. 8.10) which are accelerated by the applied electric voltage. The cathode is supplied with a tantalum cup which focusses the electrons which are emitted along the normals to the cathode surface. The target is a plate made of tungsten, platinum or some other heavy metal pressed into the anode (anode mirror) which is made of red copper for better heat removal. Impinging the surface of the target, the electrons are decelerated and produce X-rays. The voltage applied between the cathode and anode reaches several tens of kilovolts. The X-ray tube is thoroughly evacuated for the electrons to be able to reach the target without hindrance. Usually the anode is cooled with water.



**Fig. 302.**

Modern X-ray tube; the cathode circuit is not shown.



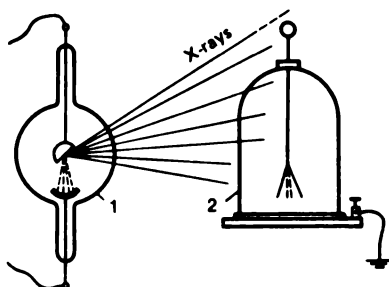


Fig. 303.

Ionization effect of X-rays: 1 — X-ray tube, 2 — electroscope. The experiment can be carried out both with a positively and a negatively charged electroscope. The ions of both signs are produced in air by X-rays.

Acting on gases, X-rays may cause their ionization (see Vol. 2, Sec. 8.2). For example, if we bring a charged electroscope close to an X-ray tube, it will soon get discharged if the tube is switched on (Fig. 303). The reason behind the leakage of charge from the electroscope is the ionization of the surrounding air by X-ray so that it becomes a conductor. The ionizing action of X-rays is also used for their detecting and recording.

### 17.6. Origination and Nature of X-rays

The method of obtaining X-rays clearly indicates that they are associated with stoppage (braking) of fast electrons. A flying electron is surrounded by electric and magnetic fields since the moving electron is a current. The stoppage (braking) of an electron involves a change in the magnetic field surrounding it, while a change in the magnetic or electric field causes (see Sec. 6.1) the radiation of electromagnetic waves. It is just these waves that are observed in the form of X-rays.

Such an idea about X-rays has already been formulated by Roentgen (although other researchers supported this idea more persistently). In order to establish the wave nature of X-rays, it was necessary to stage experiments on their interference or diffraction. However, the realization of such experiments turned out to be a complicated problem which was solved

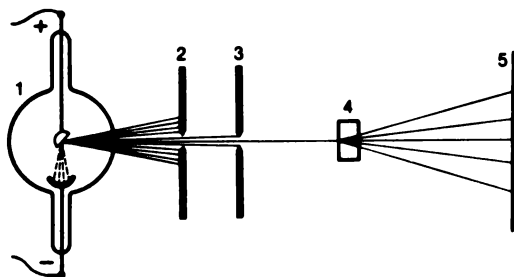
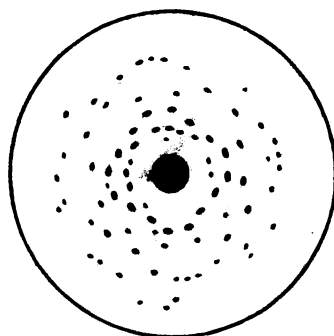


Fig. 304.

Schematic diagram of the first experiments on diffraction of X-rays: 1 — X-ray tube, 2, 3 — lead diaphragms isolating a narrow beam of X-rays, 4 — crystal by which they are diffracted, 5 — photographic plate.



**Fig. 305.**

Photograph of the diffraction pattern of X-rays in a zinc-blende crystal.

only in 1912 when the German physicist Max Laue (1879-1960) proposed to use as a diffraction grating a natural crystal in which atoms are arranged in a strict order at a distance of the order of  $10^{-10}$  m from one another (see Vol. 1, Sec. 15.4).

The diagram of experiment carried out by W. Friedrich, P. Knipping and Laue was as follows. A narrow beam of X-rays separated with the help of lead diaphragms 2 and 3 (Fig. 304) was incident on crystal 4. The image of the trace of the beam was formed on the photographic plate 5. In the absence of the crystal, the image on the plate had the form of a dark spot, viz. the trace of the beam transmitted by the diaphragm. When the crystal was arranged on the way of the beam, a complex pattern was observed on the plate (Fig. 305) which was the result of diffraction of X-rays by the crystal lattice. The obtained pattern not only provided a direct proof of the wave nature of X-rays but also allowed to draw important conclusions about the structure of the crystal, which determined the form of the diffraction pattern observed. At present, the application of X-rays for investigating the structure of crystalline and other bodies has become of high practical and scientific values.

Further modifications made it possible to determine the wavelength of X-rays with the help of thorough experiments.<sup>6</sup> It turned out that the radiation emitted by an ordinary X-ray tube, like white light, contains different wavelength with average values ranging from hundredth to tenth of a nanometre depending on the voltage between the cathode and anode of the tube. Later, X-rays with wavelengths of several tens of nanometers were obtained, i.e. the waves which are longer than the shortest of the known ultraviolet waves. Very short waves (with a wavelength of thousandth and smaller fractions of a nanometre) were also obtained and observed.

Measuring the wavelength of X-rays, one can establish that the shorter the waves, the stronger they are absorbed. Roentgen called weakly absorbed rays *hard* rays. Thus, an increase in the *hardness*<sup>7</sup> corresponds to a decrease in wavelength.

### 17.7. Scale of Electromagnetic Waves

Electromagnetic waves with a wavelength smaller than 400 nm (4000 Å) we called ultraviolet waves, while those with a wavelength larger than 760 nm (7600 Å), infrared waves. Clearly, these boundaries are rather ar-

<sup>6</sup> The employment of the diffraction of X-rays at ordinary diffraction gratings (see Sec. 14.11) for an accurate measurement of wavelength was proposed much later.

<sup>7</sup> The ability of radiation to penetrate through a substance is known as the hardness of this radiation.

bitrary, and there is no sharp change in the properties as we go over from extreme violet waves to ultraviolet waves or from limiting red waves to infrared ones. Therefore, the indications of the beginning of ultraviolet or infrared waves are tentative as well as the indications of the end of the ultraviolet and infrared regions.

The analysis of these regions involves serious difficulties associated with the fact that most of materials transparent for visible light strongly observe shorter and longer waves. Still the improvement of experimental technique made it possible to obtain and investigate infrared waves having a wavelength of a few hundreds of micrometres. On the other hand, it turned out that it is possible to obtain by electrical methods radiowaves with wavelengths of hundreds of micrometres. Thus, we have a continuous transition from visible light through the infrared waves to radiowaves.

Our information about the short-wave region has also been completed, so to say, from both sides. On the one hand, the improvement of the method of operation with ultraviolet waves made it possible to approach about 5 nm (50 Å). On the other hand, the methods of obtaining and investigating X-rays (see Sec. 17.6) with a wavelength of a few tens of nanometers have been developed with time. Thus, in the short-wave electromagnetic region we have a continuous transition from visible light through ultraviolet waves to X-rays with an infinitely small wavelengths. Very short electromagnetic waves were observed in radiation of radioactive substance (the so-called gamma-radiation, see Sec. 23.1), in cosmic rays and also during the collisions of very fast electrons accelerated in accelerators (see Sec. 23.6).

The entire scale of electromagnetic waves has been already described in Sec. 6.5 (see Fig. 125).

## Chapter 18

# Speed of Light

### 18.1. First Attempts to Determine the Speed of Light

In Sec 7.1, we considered various manifestations of light, indicating that light carries an energy. The methods of measuring this energy were specified. A natural question arises as to *what the velocity of light propagation is*.

Attempts to answer this question were made long ago. As far back as 1607 Galilei tried to determine the speed of light by the following simple experiment. Let us suppose that two observers  $A$  and  $B$  (Fig. 306) are separated by distance  $l$  and supplied with identical adjusted watches. If at a

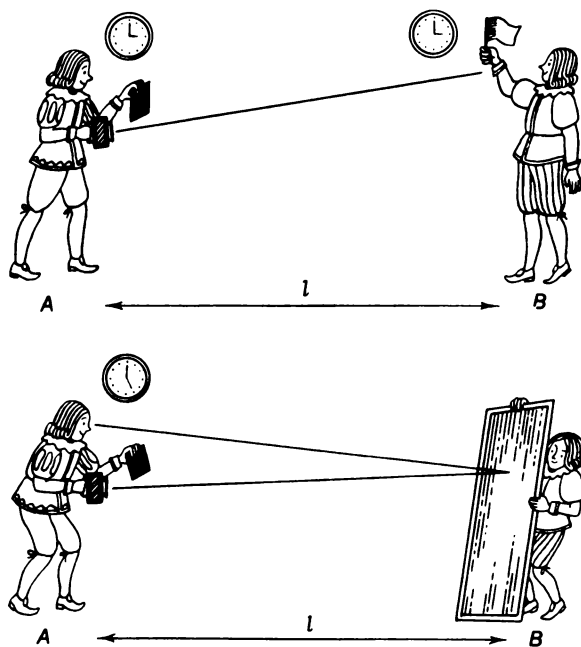


Fig. 306.

Unsuccessful attempts to determine the speed of light.

certain instant of time observer  $A$  sends a light signal (say, rapidly opens the shutter of a lantern), and observer  $B$  marks with his watch the instant when he sees this signal, it is possible to determine time  $\tau$  spent by light to cover distance  $l$ , and hence to determine the speed of light  $c = l/\tau$ .

The experiment can considerably be improved if we use a mirror instead of the second observer. The observer opening the shutter will also mark the instant of time at which the light signal reflected from the mirror returns to him, i.e. traverses distance  $2l$ . Thus, it would be possible to determine the speed of light with the help of a single time-piece. However, the Galilean experiment failed to give definite results either in the first or in the second version. Naturally, the time of emergence and arrival of the signal is registered with a certain error. On the other hand, the speed of light turned out to be so high that the time required for light to traverse comparatively short distances separating observers  $A$  and  $B$  was considerably smaller than the errors mentioned above. For this reason, the experiment which is correct in principle could not give a satisfactory result.

To improve the situation, we either had to considerably increase distance  $l$  or highly improve the accuracy of measurement of small intervals of time. Both these modifications were introduced later and led to good results.

### 18.2. Determination of the Speed of Light by Roemer

In the method proposed by the Danish astronomer Ole Roemer (1644-1710) in 1675, huge distances encountered in astronomy were used. The light signal sent from point  $A$  was an eclipse of Jupiter's satellite (say, the instant when this satellite emerges from Jupiter's shadow). An observer on the Earth recorded the time of the eclipse.

The revolution of Jupiter's satellite closest to the planet occurs in 1.75 days, i.e. eclipses follow one another quite frequently. Roemer established that eclipses are observed not completely regularly. If, for example, the moments of expected eclipses have been calculated starting from the position of the Earth  $E_1$  (Fig. 307) and observations are made from the position of the Earth half a year later,<sup>1</sup> the moment of an eclipse turns out to lag behind by nearly 16 minutes. The same calculations, however, give the correct result if observations are made when the Earth occupies position  $E_3$ , i.e. in another half a year.

Roemer gave a simple explanation to these phenomena: we must take into account the time required for the light to cover the additional distance equal to the diameter of the Earth's orbit. This additional distance is, according to modern data,  $2.99 \times 10^8$  km, the additional time being 966.4 s.

---

<sup>1</sup> The period of revolution of Jupiter is considerably longer (almost 12 times) than that of the Earth. Therefore, positions  $J_1$ ,  $J_2$  and  $J_3$  are separated by time intervals of about half a year.

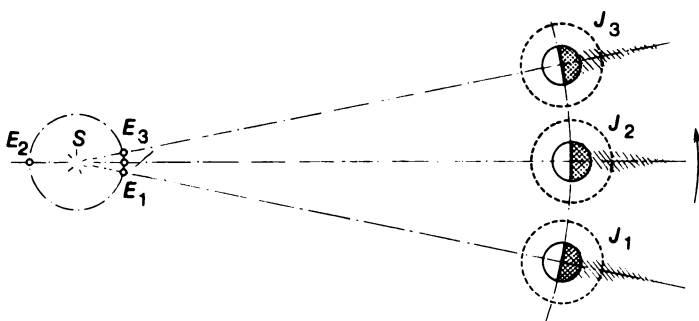


Fig. 307.

To the determination of the speed of light with the help of the Roemer method:  $J_1E_1$  — the Earth  $E_1$  is between Jupiter  $J_1$  and the Sun  $S$ ,  $J_2E_2$  — the Earth  $E_2$  and Jupiter  $J_2$  are on different sides of the Sun,  $J_3E_3$  — the next mutual arrangement of the Earth  $E_3$  and Jupiter  $J_3$ .

Hence the speed of light  $c$  is about 300 000 km/s. The value obtained by Roemer himself was 215 000 km/s.

### 18.3. Measurement of the Speed of Light by Rotating-Mirror Method

In 1862, the French physicist Jean Bernard Léon Foucault (1819-1868) developed a very accurate method of measuring the time of passage of light between two points  $A$  and  $B$ . This made it possible to measure the speed of light without resorting to extremely large distances between points  $A$  and  $B$ .

A light signal propagating in direction  $SA$  (Fig. 308) was reflected by a rotating mirror  $A$  towards a stationary mirror  $B$ . The latter had a spherical shape with a very large radius of curvature  $R$  so that its centre coincided with mirror  $A$ . Owing to such an arrangement, light propagated only along

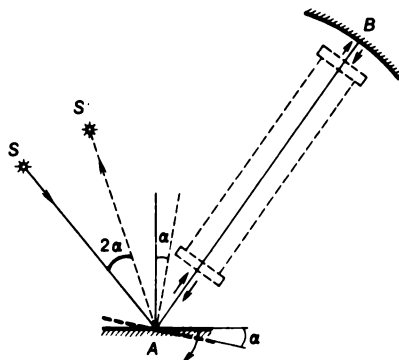


Fig. 308.

To the determination of the speed of light with the help of the rotating mirror method.

$(9.461 \times 10^{12})$  km. This value was found convenient as a unit of length for huge astronomical distances and was called a *light year*.

Along with this unit, astronomers also use parsecs. A *parsec* (viz. parallax-second) is the distance from which the radius of the Earth's orbit ( $150 \times 10^6$  km) is seen at an angle of  $1''$ . One can easily calculate that a parsec is equal to about 3.25 light years.

It has now become possible to measure independently the frequency  $\nu$  and the wavelength  $\lambda$  of a monochromatic light. Therefore, the speed of light  $c = \lambda\nu$  can be determined without kinematic measurements required for old methods.

the radius of mirror  $B$  for any position of mirror  $A$ , was incident on mirror  $B$  along the normal to its surface and, after the reflection, propagated along the radius of  $B$ , i.e. returned to mirror  $A$ . However, during time  $\tau$  required for light to traverse the distance from  $A$  to  $B$  and back (i.e. the distance of  $2R$ ), mirror  $A$  had time to rotate through a small angle  $\alpha$ , and light was reflected in direction  $AS'$  making an angle of  $2\alpha$  with direction  $SA$ . Having measured the angle  $2\alpha$  and knowing the angular velocity of mirror rotation, it is possible to determine time  $\tau$ , and hence the speed of light  $c = 2R/\tau$ .

In one of Foucault's experiments, distance  $AB = 4\text{m}$ , the mirror rotational frequency  $N = 800\text{ s}^{-1}$  and the angle of rotation  $\alpha = 27.3''$ . Consequently, for these data we have

$$\tau = \frac{\alpha}{2\pi N} = 2.7 \times 10^{-8}\text{ s and } c = \frac{2R}{\tau} = 296\,000\text{ km/s.}$$

The mean value of the speed of light obtained by Foucault was  $298\,000\text{ km/s}$ .

Having placed a water-filled tube at the route  $AB$  of light, Foucault was able to measure directly the velocity of light propagation in water. He obtained a value equal to  $4/3$  of the speed of light in air, which was in accord with Huygens' ideas (see Sec. 14.3).

The American physicist Albert Abraham Michelson (1852-1931) introduced a number of interesting modifications into the rotating-mirror method and managed to considerably improve the accuracy of determining the speed of light. According to his data (1927),  $c = 299\,796\text{ km/s}$ . In recent years, the laboratory methods of determining the speed of light have considerably been improved. They were based on independent measurements of the light wavelength and frequency. Using this method, K. Evenson with coworkers determined in 1972 the speed of light to within  $0.2\text{ m/s}$ :  $c = 299\,792\,456.2 \pm 0.2\text{ m/s}$ . These results, however, require a further confirmation. According to the resolution adopted in 1973 at the General Assembly of International Committee on Numerical Data for Science and Technology, which generalized all the known experimental results, the speed of light in vacuum was assumed to be

$$c = 299\,792\,458 \pm 1.2\text{ m/s.}$$

In practical calculations, we shall assume that the speed of light in vacuum is  $300\,000\text{ km/s}$  ( $3 \times 10^8\text{ m/s}$ ).

The speed of light, which has an enormously high value on terrestrial scale, is not so high on astronomical scale, where the time of light propagation is measured in very large figures. For example, light travels about 8 minutes from the Sun to the Earth, and from the nearest star, about 4 years. The distance traversed by light during a sidereal year is about  $10^{13}$



## Chapter 19

# Dispersion of Light and Colours of Bodies

### 19.1. State-of-the-art in Chromatography Before Newton's Studies

The question about the reason behind different colours of bodies has naturally attracted the attention of researchers since old times. A very large number of observations in everyday life and in scientific experiments were available to scientists. However, before Newton published his works (it was about 1666), there had been much confusion in this problem. Colour was believed to be a property of a body itself, although thorough observations showed that considerable variations of the colour of the same body often take place depending on the time of day or illumination conditions. According to another opinion, different colours were obtained as a result of “mixing” light and dark shades. In other words, two essentially different concepts, colour and illuminance, were confused. From ancient times, people observed beautiful (rainbow) colours. Even at that time it was known that a rainbow is formed by the passage of light through rain drops. For example, the French mathematician and physicist René Descartes (1596-1650) observed an artificial rainbow in sprinklings of fountains and made experiments on producing rainbows with the help of glass spheres filled with water. In 1637, Descartes explained the shape and angular size of a rainbow of the firmament. However, the reason behind rainbow colours and their sequences remained unclear.

Similarly, it was well known how colours play in cut diamonds and even in glass prisms. In eastern countries, e.g. in China, decorations in the form of glass prisms producing rainbow patterns were in great demand. These Chinese toys were also known in Europe. Nevertheless, nobody tried to correlate these numerous and diversified phenomena. The relationship between beautiful colours of the rainbow playing in the sky and colours of bodies was established only in the brilliant studies of Newton.

### 19.2. Main Discovery of Newton in Optics

Newton started to study colours observed during refraction of light in connection with the attempts made to improve telescopes. Trying to obtain lenses of as high quality as possible, Newton came to the conclusion that

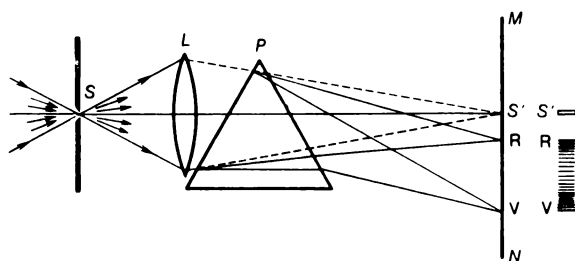


Fig. 309.

Schematic diagram of Newton's fundamental experiment on dispersion of light. The distance from the screen to the prism is large enough to distinguish between separate colour fringes.

the main drawback of the image is coloured fringe. It is well known that this circumstance motivated a changeover to telescopes with mirrors (reflector telescopes) (see Sec. 12.10). Investigating colours appearing as a result of refraction, Newton made his greatest optical discoveries.

The essence of Newton's discovery can be explained with the following experiments (Fig. 309). Light from a lantern illuminates a narrow aperture (slit)  $S$ . Lens  $L$  is used to obtain the image of the slit on screen  $MN$  in the form of a short white rectangle  $S'$ . Having located prism  $P$  whose edge is parallel to the slit on the path of the rays, we shall see that the image of the slit will be shifted and transformed into a coloured strip in which the transitions from red to violet colours are similar to those observed in a rainbow. This coloured image of the slit was called a *spectrum*<sup>1</sup> by Newton (Fig. 310).

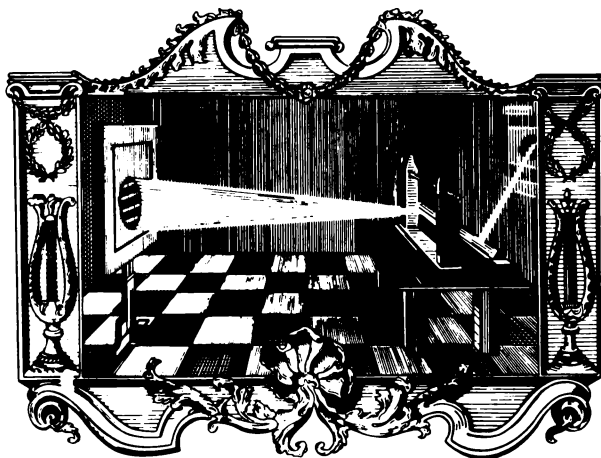


Fig. 310.

Newton's experiment on dispersion of sunlight (the drawing by Academician Kraft, kept in the cabinet of curiosities of the Academy of Sciences, 18th century).

<sup>1</sup> The Latin *spectrum* means appearance, image.

If we cover the slit with a coloured glass, i.e. direct a coloured beam on the prism instead of white, the image of the slit will be transformed to a coloured rectangle lying in the corresponding region of the spectrum. In other words, light will be deflected from the initial image  $S'$  through different angles depending on the colour. This experiment shows that *rays of different colours are refracted by a prism differently*.

This important conclusion was verified by Newton in many experiments. The most important of them consisted in determining the refractive index for rays of different colours forming a spectrum. For this purpose, a slit was cut in screen  $MN$  (see Fig. 309) on which a spectrum was obtained. By shifting the screen, it was possible to single out a narrow beam of a certain colour and let it through the slit. This way of isolating monochromatic rays is better than the one using a coloured glass. Experiments revealed that after refraction in the second prism, such an isolated beam is not stretched into a strip any longer. The beam has a certain refractive index whose value depends on the colour of the isolated beam.

### 19.3. Interpretation of Newton's Observations

Experiments described above show that for a narrow coloured beam isolated from the spectrum the refractive index has a quite definite value, while the refraction of white light can be characterized by a certain value of refractive index only approximately. Comparing observations of this type, Newton drew the conclusion that there are simple colours that cannot be decomposed by a prism and complex colours that are combinations of simple colours corresponding to different refractive indices. In particular, solar light is such a combination of colours which is decomposed by a prism to produce the spectral image of a slit.

Thus, the fundamental Newton's experiments contained two important discoveries: (1) *different colours are characterized by different refractive indices in a given substance (dispersion)*<sup>2</sup>, (2) *white light is a combination of simple colours*.

It is now known that different colours correspond to different wavelengths. Therefore, Newton's first discovery can be formulated as follows: *the refractive index of a substance depends on the wavelength of light*. Normally, it increases with decreasing wavelength.

The formulation of Newton's first discovery remains unchanged in modern physics. As to the second statement, it should be mentioned that the question about the nature of white light is very complicated. This problem is beyond the scope of this book.

<sup>2</sup> Dispersion is derived from the Latin *dispersus* meaning to strew, scatter + ion. It would be more correct to call the phenomenon observed by Newton the *dispersion of refractive index* since other optical quantities also exhibit the dependence of wavelength (dispersion).

However, for a very large number of practical problems, white light can be replaced by a combination of specially selected simple (monochromatic) colours, i.e. be treated as the mixture of these colours.

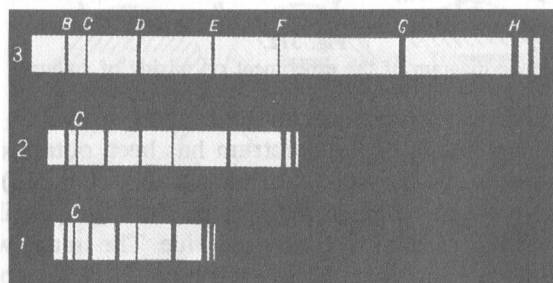
The discovery of the decomposition of white light into colours as a result of refraction has made it possible to explain the formation of a rainbow and similar meteorological phenomena. Refraction of light in rain drops or ice crystals floating in the atmosphere is accompanied by the decomposition of sunlight as a result of dispersion in water or ice. Calculating the direction of refracted rays in the case of spherical water drops, we obtain the pattern of distribution of coloured arcs, which exactly corresponds to those observed in a rainbow. Similarly, an analysis of refraction of light in ice crystals allows one to explain the presence of the solar and lunar halos in winter, the formation of so-called false suns, poles, etc.

#### 19.4. Dispersion of Refractive Indices for Different Materials

The measurements of refractive index as a function of wavelength for different substances show that dispersion of different materials can differ considerably. Table 9 represents, by way of an example, the values of refrac-

**Table 9.** Refractive Indices of Various Substances  
as Functions of Wavelength

| Wavelength $\lambda$<br>in nm<br>(colour) | Refractive index |                      |                |        |
|-------------------------------------------|------------------|----------------------|----------------|--------|
|                                           | crown<br>glass   | carbon<br>disulphide | flint<br>glass | water  |
| 656.3 (red)                               | 1.6444           | 1.5145               | 1.6219         | 1.3311 |
| 589.3 (yellow)                            | 1.6499           | 1.5170               | 1.6308         | 1.3330 |
| 486.1 (blue)                              | 1.6657           | 1.5230               | 1.6799         | 1.3371 |
| 404.7 (violet)                            | 1.6852           | 1.5318               | 1.6990         | 1.3428 |



**Fig. 311.**

Comparison of dispersion for different substances: 1 — water, 2 — light crown glass, 3 — heavy flint glass. On the dark lines in the spectrum see Sec. 20.7.

tive index as a function of wavelength for two grades of glass and two different liquids.

Figure 311 shows the spectrum of solar radiation which could be obtained with the help of prisms of the same shape, made of the substances represented in the table.

The difference in dispersion for different grades of glass makes it possible to correct chromatic aberration as was mentioned in Sec. 11.8.

### 19.5. Complementary Colours

As was mentioned in Sec. 19.2, Newton's fundamental experiment consisted in the decomposition of white light into spectrum. It is natural to expect that if we mix all the colours of the obtained spectrum, we shall get white light again. Such experiments were also carried out by Newton. Spectral colours can be mixed, for example, as follows. We direct a parallel beam of white light on prism  $P$  (Fig. 312). Diaphragm  $D$  is placed at the exit face of the prism, and lens  $L$ , behind the prism. A coloured strip  $rv$  will be obtained in the principal focal plane  $MN$  of the lens where parallel beams of different colours converge since rays of different colours are incident on the lens at different angles and hence are converged at different points of the focal plane. But these coloured beams of rays passing through diaphragm  $D$  in different directions will form, owing to lens  $L$ , the image of diaphragm  $D$  in the form of white circle in plane  $AB$ . At each point of the image, all the rays which constituted the white beam incident on the prism are mixed.

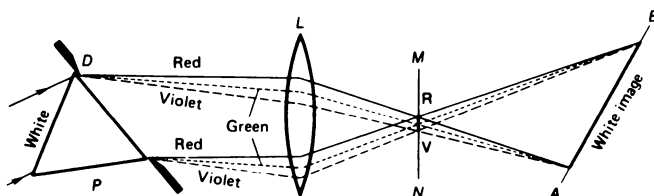
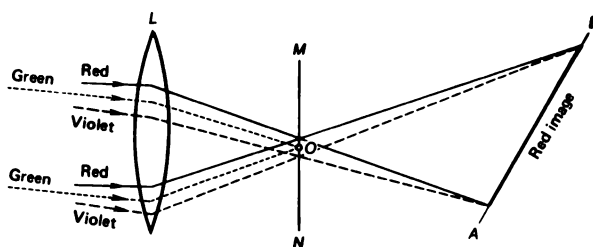


Fig. 312.

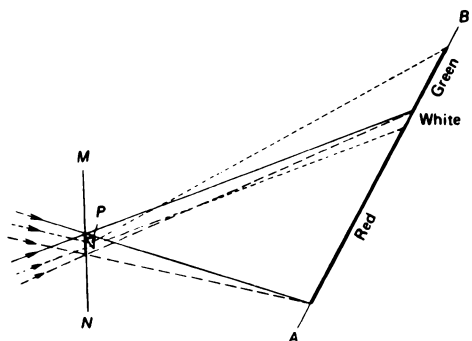
Schematic diagram of the experiment on mixing of colours.

Let us now place an opaque narrow body (say, pencil  $O$ ) into plane  $MN$  where the sharp image of the spectrum has been obtained so that it covers a certain region of the spectrum (say, green) (Fig. 313). Then the image turns out to be coloured in red. Let us shift the pencil so that it covers some other rays in the spectrum, say, blue. The image will become yellow. By moving the pencil parallel to itself along  $MN$ , i.e. consecutively covering one colour after another, we shall change the colour of the image since, for each position of the pencil, not all colours of rays constituting white light will participate in image formation, but only a part of them.

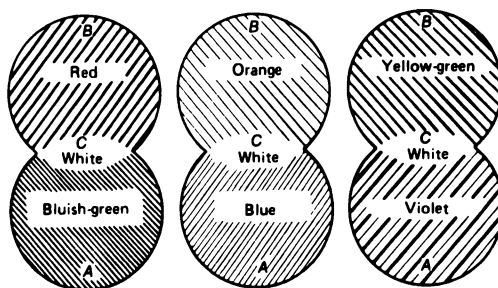
**Fig. 313.**

Pencil *O* cuts out a green part of the spectrum.

This experiment becomes even more visual if a part of the rays is deflected by arranging a mirror or a prism on their way (Fig. 314).

**Fig. 314.**

Prism *P* deflects a green part of the spectrum.

**Fig. 315.**

Overlapping of images in complementary colours, obtained by the method represented schematically in Fig. 314.

In this case, we shall obtain on screen  $AB$  two images arranged close to each other. One of them is formed by deflected rays, while the other by the remaining rays of the spectrum. The images will be coloured. If the angle of deflection is chosen so that coloured images partially overlap, the common part of the image will be illuminated by all the rays of the spectrum and will be white.

Thus, the pattern will be similar to that shown in Fig. 315. Hatched parts  $A$  and  $B$  have different colours, while part  $C$  is white. The colours of regions  $A$  and  $B$  are called *complementary* since they complement each other to white.

Modifying the above experiments, one can choose a large number of combinations of complementary colours. Some of them are compiled in Table 10.

**Table 10.** Complementary Colours

| Isolated part of spectrum           | bluish-green | green  | orange | red          | yellow | yellow-green |
|-------------------------------------|--------------|--------|--------|--------------|--------|--------------|
| Colour of mixture of remaining rays | red          | purple | blue   | bluish-green | indigo | violet       |

Complementary colours can also be obtained by using specially selected coloured glasses. If the glasses are selected correctly, two coloured images obtained with their help and partially overlapping produce a pattern similar to that shown in Fig. 315. Two complementary colours taken together may not represent the entire spectrum. For example, a narrow region of red light satisfactorily complements the corresponding region of green colour. However, the most perfect complementary colours are those obtained by dividing the spectrum of white light into two parts.

### 19.6. Spectral Composition of Light Emitted by Various Sources

In Newton's experiments, it was established that the sunlight has a complex composition. Similarly, i.e. by analyzing the composition of light with the help of a prism, one can see that the light from most of other sources (an incandescent lamp, arc lamp, etc.) is of the same nature. Comparing the spectra of these luminous bodies, we can reveal that the relative regions of the spectra have different brightnesses, i.e. the energy distribution is different in different spectra. Still more reliable data can be obtained when spectra are investigated by means of a thermocouple (see Sec. 17.1).

Although for conventional light sources, these differences in spectra are insignificant, they can easily be observed. Even without a spectroscopic instrument, the eye reveals the differences in the properties of white light emitted by these sources. For instance, the light from a candle appears as yellow or even reddish in comparison with the light from an incandescent lamp, the latter giving more yellow colour than the Sun.

The differences become even more pronounced if a light source is a gas-filled tube glowing under the action of electric discharge rather than a red-hot body. Such tubes are now used for illuminating streets or as advertising lights. Some of these gas-discharge lamps emit yellow light (sodium lamps) or red light (neon lamps), while others have a white glow (mercury lamps) which differs in shade from the sunlight. A spectral analysis of light from such sources reveals that their spectra contain only individual more or less narrow coloured regions.

At present, gas-discharge lamps are designed whose light has a spectral composition close to that of the sunlight. The lamps are called fluorescent lamps (see Sec. 21.3).

If we analyze the light emitted by the Sun or an arc lamp, filtered through a coloured glass, it will be found to differ considerably from the initial radiation. The eye perceives this light as coloured, and its spectral decomposition reveals that some (more or less significant) parts of its spectrum are either missing or very weak.

### 19.7. Light and Colours of Bodies

Experiments described in Sec. 19.6 show that light causing the sensation of a certain colour in the eye has a more or less complex spectral composition. It turns out that the eye is an imperfect instrument for an analysis of light so that the rays of different spectral compositions may sometimes cause nearly similar sensation. Nevertheless, it is the eye which gives us the knowledge of the diversity of colours in the surrounding world.

Light from the source is directed to the eye of an observer in comparatively rare cases. More often, light has first to pass through various bodies, being refracted and partially absorbed by them, or being reflected to a certain extent from their surface.

Thus, the spectral composition of light reaching the eye may be considerably changed due to the processes of reflection, absorption, etc. described above. In the majority of cases, all such processes just cause an attenuation of certain spectral regions and may even completely eliminate some of them without adding wavelengths which were not contained in it to the light emitted by a source. However, the latter processes are also observed sometimes (as, for example, in fluorescence).



### 19.8. Absorption, Reflection and Transmission Coefficients

Different objects illuminated by the same light source (say, the Sun) may have quite different colours in spite of the fact that they all are illuminated by light of the same spectral composition. The main role in such effects is played by reflection and transmission of light. As was mentioned above, a luminous flux incident on a body is partially reflected (scattered), partially transmitted and partially absorbed by it. The fraction of the luminous flux participating in each of these processes is determined by the relevant coefficients: of reflection  $\rho$ , transmission  $\tau$  and absorption  $\alpha$  (see Sec. 8.9).

Each of these coefficients ( $\alpha$ ,  $\rho$  and  $\tau$ ) may depend on the wavelength (colour), due to which various effects are observed during illumination of a body. It can easily be seen that a body having, for example, a large transmission coefficient and a small reflection coefficient for the red light, and vice versa, for the green light will appear red in transmitted light and green in reflected light. Such properties are exhibited, for instance, by chlorophyll, viz. the green substance contained in the leaves of plants and responsible for their green colouring. A chlorophyll solution (extract) in alcohol appears red in transmitted light and green in reflected light.

Bodies characterized by large absorption and small reflection and transmission for all types of rays are black nontransparent bodies (like carbon black). For a very white opaque body (like magnesium oxide), coefficient  $\rho$  is close to unity for all wavelengths, while coefficients  $\alpha$  and  $\tau$  are very small. A completely transparent glass has small coefficients  $\rho$  and  $\alpha$  and coefficient  $\tau$  close to unity for all wavelengths. On the contrary, a coloured glass has coefficients  $\tau$  and  $\rho$  nearly equal to zero for certain wavelengths and, accordingly, the value of coefficient  $\alpha$  close to unity. The difference in the values of coefficients  $\alpha$ ,  $\tau$  and  $\rho$  and their dependence on colour (wavelength) explains a rich diversity of colours and shades of different bodies.

### 19.9. Coloured Bodies Illuminated by White Light

Painted bodies appear to be coloured when illuminated by white light. If the layer of paint is thick enough, the colour of a body is determined by it and does not depend on the properties of deeper layers under the paint. Usually, paint forms fine grains nonuniformly scattering light and immersed in a transparent binding material like oil. The coefficients  $\alpha$ ,  $\rho$  and  $\tau$  of these grains determine the properties of a paint.

The action of a paint is shown schematically in Fig. 316. The uppermost layer reflects almost equally all rays, i.e. scatters white light. Its fraction is insignificant (about 5%). The remaining 95% of light penetrates into the bulk of the paint layer and emerges from it having been scattered by

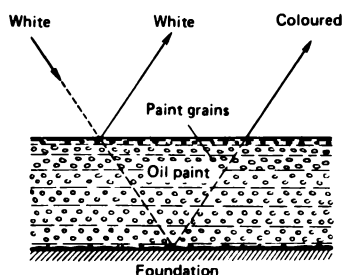


Fig. 316.

A paint layer illuminated by white light.

paint grains. In the process, a part of light is absorbed in the grains so that some of spectral regions are absorbed to a larger or smaller extent depending on the colour of the paint. A part of light penetrating still deeper is scattered in the lower-lying layers of grains, and so on. As a result, the body illuminated by white light will have the colour determined by the values of coefficients  $\alpha$ ,  $\tau$  and  $\varrho$  for the grains of the paint coating the body.

The paints absorbing the incident light in a very thin layer are known as *covering paints*. The paints whose action is determined by a large number of grain layers are known as *scumbling paints*. The latter give good effects by mixing a few types of coloured grains (trituration on a palette). As a result, various colour effects can be obtained. It is interesting to note that by mixing the scumbling paints corresponding to complementary colours, we must obtain very dark shades. Indeed, let us suppose that red and green grains are mixed in a paint. The light scattered by red grains will be absorbed by green ones, and vice versa, so that light will be absorbed almost completely by the layer of paint. Thus, the mixing of paints gives completely different results than the mixing of light of the corresponding colours. This should be borne in mind by an artist mixing paints.

### 19.10. Coloured Bodies Illuminated by Coloured Light

All what was said above pertains to illumination by white light. If the spectral composition of incident light differs significantly from daylight, the effects of illumination may be completely different. Bright coloured regions look dark if in the incident light there are no wavelengths for which these regions have just a large reflection coefficient. Even a transition from daylight illumination to artificial light at night may considerably change the relation of shades. In the daylight, the relative fraction of yellow, green and blue rays is much larger than in artificial light. For this reason, yellow and green cloths look much dimmer at night than in daytime, while a blue cloth in daylight seems to be almost black at night under a lamp. This should be taken into consideration by artists and decorators choosing colours for a performance or for a parade held during daytime in the open.

In many branches of light industry, where the correct estimate of shades, say, of yarn, is very important, nightwork is undesirable or even impossible. Under these conditions, it is expedient to use fluorescent lamps, i.e. those whose spectral composition would be very close to that of daylight illumination (see Sec. 21.6).

### 19.11. Masking and Unmasking

Even with a bright illumination, we are unable to distinguish between bodies whose colour does not differ from that of a background, i.e. the bodies for which coefficient  $\rho$  has practically the same value for all wavelengths as that of the background. For example, it is difficult to distinguish between animals with white fur or people in white clothes in a snow field. This is used in army for camouflaging troops and military objects. In nature, many animals have acquired protecting colours of skin in the process of natural selection (protective colouration).

It is clear from what was said above that the most perfect masking involved a selection of colouration whose reflection coefficient  $\rho$  for all wavelengths has the same value as that of the background. It is difficult to attain such an effect in actual practice, and one has often to limit oneself to the selection of close reflection coefficients for radiation that plays the most important role in daylight illumination and visual observation. This is mainly the yellow-green part of the spectrum to which the eye is most sensitive and which is most extensively represented in the solar (daylight) spectrum. If, however, the objects masked in this way are not observed by the eye but are photographed, masking may lose its significance. Indeed, a photographic plate is mostly affected by the violet and ultraviolet radiation. Therefore, if the reflection coefficient of the object for this spectral region differs significantly from that of the background, this defect of masking remains unnoticed in visual observation but is clearly seen on a photograph. The defects of masking also become obvious if observations are made through a light filter which practically absorbs the wavelengths for which the masking was mainly designed, say, through a blue filter. In spite of a considerable reduction in the brightness of the entire pattern observed through such a filter, it may show details which were hidden during the observation in white light. A combination of a filter and photography may produce an especially strong effect. For this reason, while choosing masking colours, special attention should be paid to determining  $\rho$  for a sufficiently broad spectral region, including the infrared and ultraviolet regions.

Light filters are sometimes used to ensure the correct recording of illuminance in photography. Since the sensitivity maxima of the eye and of a photographic plate lie in different regions (in the yellow-green for the eye and in the blue-violet for the photographic plate), the visual and photographic sensations may be quite different. The figure of a girl dressed in

a yellow blouse and a violet skirt is perceived by the eye as light in its upper part and dark in its lower part. However, on a photograph she may turn out to be dressed in a dark blouse and light skirt. If, however, we arrange a yellow filter in front of the camera objective, it will change the relation between the illuminances of the blouse and the skirt so that it will become closer to the visual sensation. If, moreover, we use a photographic film with an elevated sensitivity to long waves (orthochromatic film) in comparison with an ordinary film, we can attain a rather correct reproduction of the illuminance of the figure.

### 19.12. Colour Saturation

In addition to colour name (red, yellow, blue, etc.), we often distinguish colours by their *saturation*, i.e. the purity of a shade and the absence of whitishness. Spectral colours may serve as examples of deep, or saturated, colours. They represent a narrow region of wavelengths without any admixture of other colours. On the other hand, the colours of cloths and paints covering objects are normally less saturated and more or less whitish. The reason behind this is that the reflection coefficient for most dyes is not zero for any wavelength. Thus, when a dyed cloth is illuminated by white light, we observe in the scattered light mainly a single spectral region (say, red) containing, however, a considerable admixture of other wavelengths which give together white light. But if such a light scattered by the cloth and containing mainly one colour (say, red) is directed not right to the eye but is made to be reflected for the second time from the same cloth, the fraction of dominating colour will increase considerably in comparison with other colours, and the cloth will appear less whitish. A multiple repetition of this process (Fig. 317) may lead to a sufficiently saturated colour.

If we denote by  $I$  the intensity of an incident light having a certain wavelength and by  $q$  the reflection coefficient for the same wavelength, the intensity of light after the first reflection will be  $Iq$ , after the double reflection it will be  $Iq^2$ , after the triple reflection,  $Iq^3$ , and so on. This means that if the value of  $q$  for some narrow spectral region is, for example, 0.7, and for the remaining regions it is 0.1, the admixture of white light after a single reflection

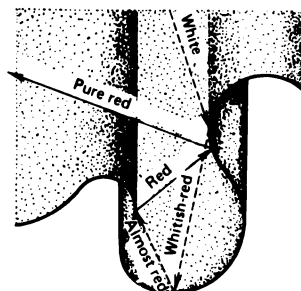


Fig. 317.

Obtaining saturated colour in reflection from red curtains.

will be  $1/7$ , i.e. about 15%, after the double reflection it will be  $1/49$ , i.e. about 2%, while after the triple reflection it will be only  $1/343$ , i.e. less than 0.3%. Such a light can be regarded as completely saturated.

The phenomenon described above explains the saturated colours of velvet cloths, folds of curtains or flutters of banners. In all these cases, there are numerous depressions (velvet) or folds of a dyed cloth. White light incident on them has undergone multiple reflection before it reaches the eye of an observer. Naturally, the cloth appears to be darker than, for example, a smooth stretched strip of a satin cloth. The colour saturation increases drastically, and the cloth looks much more beautiful.

In Sec. 19.9 it was mentioned that the surface layer of any paint always scatters white light. This deteriorates the colour saturation of paintings. For this reason, oil paintings are usually covered with a layer of varnish. Covering all uneven regions of the paint, the varnish layer produces a mirror surface of the painting. White light is not scattered by such a surface in all directions but is rather reflected in a certain direction. Naturally, if such a painting is viewed from a wrongly chosen position, its perception will be deteriorated (reflection). But if we view the painting from other positions, white light will not propagate in these directions owing to the varnish coating, and the colour saturation of the painting will considerably be improved.

### 19.13. Colour of the Sky and Dawns

The change in the spectral composition of light reflected or scattered by the surfaces of bodies is due to *selective absorption* and *reflection*, which is expressed in the dependence of coefficients  $\alpha$  and  $\rho$  on wavelength.

There is one more phenomenon in nature that plays a major role in the change in the spectral composition of solar radiation. Light reaching an observer from the regions of cloudless sky far from the Sun is characterized by a saturated blue or even indigo shade. Undoubtedly, this is the sunlight scattered in the bulk of the atmosphere and for this reason reaching the observer from all directions differing from the direction to the Sun. Figure 318 explains the origin of the scattered light of the sky.

Theoretical investigations and experiments showed that such a scattering is due to the molecular structure of air. Even when air is completely free of dust, it scatters the solar radiation. The spectrum of the light scattered by the atmosphere differs considerably from the spectrum of direct solar radiation. In solar radiation, the maximum of energy corresponds to the yellow-green spectral region, while in the scattered light of the sky, this maximum is shifted to the blue region. The reason behind this is that *short light waves are scattered considerably stronger than long waves*. According to the calculations made by the English mathematician and physi-

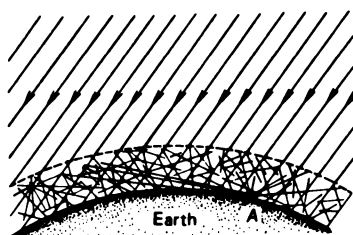


Fig. 318.

To the explanation of the colour of the sky (the sunlight is scattered by the atmosphere). Direct sunlight, as well as the light scattered in the atmosphere, reach the surface of the Earth (say, point *A*). The colour of the scattered light determines the colour of the sky.

cist John William Strutt (Lord Rayleigh) (1842-1919) confirmed by measurements, the intensity of scattered light is inversely proportional to the fourth power of the wavelength if scattering particles are small in comparison with the wavelength. Consequently, violet rays are scattered nine times as strongly as red rays. For this reason, the yellowish light from the Sun is converted upon scattering in the atmosphere into the blue colour of the sky. This is the case when the atmospheric air is clean (as in mountains or above the ocean). The presence of comparatively large dust particles in air (in cities) adds the light scattered by dust particles to the scattered blue light, i.e. an almost unchanged solar radiation. Due to this admixture, the colour of the sky in such regions is more whitish.

The predominant scattering of short waves explains why the direct solar radiation reaching the Earth is more yellow than light observed at a high altitude. On its way to the Earth, solar radiation is partially scattered. Since short waves are scattered to a larger extent, the light reaching the Earth turns out to be richer in the long-wave part of the spectrum. This phenomenon is the most pronounced during sunrise or sunset (the same is true for the Moon) when direct rays have to traverse through a much thicker layer of atmospheric air (Fig. 319). As a result, the rising or setting Sun (Moon)

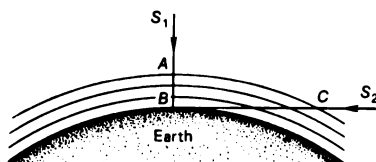
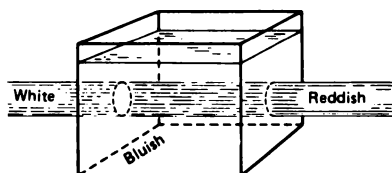


Fig. 319.

To the explanation of the red colour of the Moon and the Sun at moonrise (dawn) and at moonset (dusk):  $S_1$  — a celestial body is at zenith, which corresponds to the shortest path length ( $AB$ ) in the atmosphere and  $S_2$  — a celestial body is at the horizon, which corresponds to the longest path length ( $CB$ ) in the atmosphere.



**Fig. 320.**

Dissipation of light by a turbid liquid: incident light is white, dissipated light is bluish, while transmitted light is reddish.

has a copper-coloured or even a reddish tinge. If the atmospheric air contains very small (in comparison with the wavelength) dust particles or drops of water (mist), the scattering caused by their presence also follows a law similar to the Rayleigh scattering law, i.e. short waves are scattered to a larger extent. In such cases, the rising and setting Sun may be perfectly red. The clouds in the sky are also red in colour. This explains the beautiful crimson shades at dawn and dusk.

The change in colour during scattering can be observed if we pass the light from a lantern through a vessel containing an opaque liquid (Fig. 320), i.e. the liquid containing small suspended particles (e.g. water with a few drops of milk). The rays propagating sideways (scattered rays) have more saturated blue colour than the direct light from the lantern. If the thickness of the opaque liquid is large enough, the light passing through the vessel loses such a large part of its short-wavelength rays (blue and violet) that it becomes orange or even red.

In 1883, a huge eruption of volcano in the Krakatau Island was observed, as a result of which half the island was destroyed and an enormous amount of dust was ejected to the atmosphere. This dust carried by air flows over very large distances contaminated the atmosphere during several years and caused especially red dawns.

## Chapter 20

# Spectra and Spectral Regularities

### 20.1. Spectroscopic Instrumentation

The glow of bodies is associated with processes occurring in atoms and molecules. For this reason, the investigation of glow has become a powerful method for explaining the structure of atoms and molecules.

Considerable difference in the luminescence mode is observed by analyzing the spectra of luminous bodies. Such spectra are obtained with the help of diffraction grating or, which is more often, a prism. The principle of obtaining spectra with the help of a prism is explained in Sec. 19.2. In order to obtain a clear spectrum, i.e. such that spectral regions can easily be distinguished, a spectroscopic instrument is made more complicated than the one described in Sec. 19.2. The schematic diagram of a spectrograph is shown in Fig. 321.

The left-hand part of the instrument is *collimator*  $SL_1$  consisting of a narrow slit  $S$  arranged in the principal focal plane of objective  $L_1$ . Due to such an arrangement, the light incident on the slit emerges from the collimator in a parallel beam and is incident on a prism. It emerges from the prism also in a parallel beam. However, since the rays of different wavelengths (different colours) are deflected by the prism through different angles (dispersion), parallel beams having different directions emerge from the prism. As a result, light is converged by the second objective  $L_2$  at different points of its focal plane  $MN$ . Thus, in this plane we obtain the images

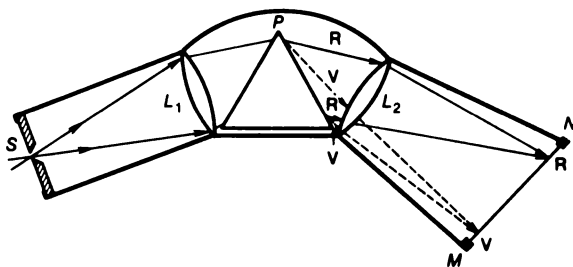


Fig. 321.

Schematic diagram of a spectrograph:  $S$  — slit,  $L_1$  — objective of the collimator,  $P$  — prism,  $L_2$  — objective of the camera and  $MN$  — opaque glass or photographic plate.



of slit  $S$  but such that the images corresponding to different wavelengths are formed in different regions of plane  $MN$ . Having placed an opaque glass or a photographic plate in this plane, we obtain a clear image of the spectrum on it. If the light incident on slit  $S$  is a combination of several monochromatic beams, the spectrum has the form of separate images of the slit for different wavelengths, i.e. individual narrow lines separated by dark gaps. If white light is incident on the slit, all the individual images of the slit merge into a coloured band.

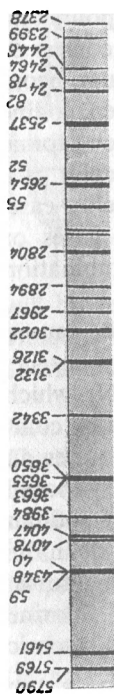
The instruments in which spectra are formed on a photographic plate are known as *spectrographs*. Sometimes, a telescope is used instead of chamber  $L_2MN$  and spectra are observed by the eye. In such a case, the spectroscopic instrument is called a *spectroscope*. The prism is made of glass having considerable dispersion (quartz, fluorite or rock salt) if a spectrograph is designed for investigating the ultraviolet or infrared region of the spectra. Objectives are also made of appropriate materials.

## 20.2. Types of Emission Spectra

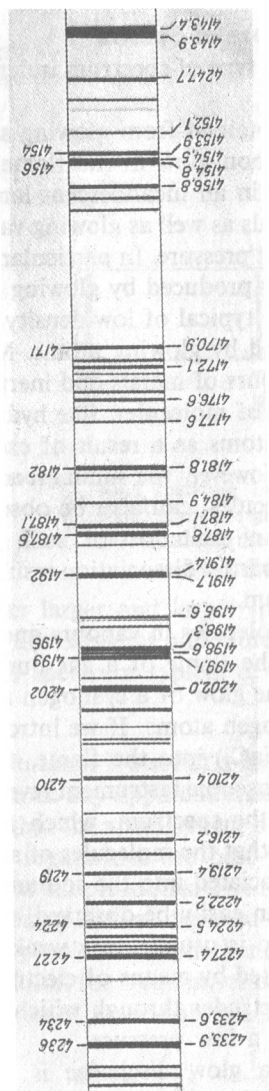
Having directed the light from the Sun, an incandescent lamp, candle, etc. on the spectrograph slit, we obtain spectra in the form of a continuous band in which all wavelengths are represented in a sequence. Such spectra are known as *continuous spectra*.

A spectrum has a different form if a source of light is a glowing gas. If we direct, for example, the light from a gas-discharge mercury-vapour tube on a spectrograph, the observed spectrum will have the form shown in Fig. 322. It consists of separate sharp lines representing the image of the spectrograph slit in individual wavelengths. Each line is essentially a narrow spectral interval embracing a set of wavelengths. This interval, however, is so narrow that it can practically be assumed to correspond to a single wavelength. The spectrum of mercury shown in Fig. 322 by way of an example is typical of glowing gases or vapours. Such spectra are called the *line spectra*. Various vapours and gases may give spectra differing in position of spectral lines (i.e. in wavelength) as well as in the number and distribution of lines over a spectrum. The spectrum of mercury vapour is not rich in lines. On the contrary, the spectrum of iron vapour, for example, contains a few thousand individual spectral lines (Fig. 323) distributed over the visible and ultraviolet spectral regions.

An analysis of spectra of vapours and gases revealed that there are also spectra consisting of individual bands separated by dark gaps. A more thorough investigation showed that some of these bands consist of a large number of separate lines, while others are indeed continuous bands. Such spectra are known as *band spectra*. Figure 324 represents an example of such a spectrum observed from glowing iodine vapour.



**Fig. 322.**  
Spectrum of mercury vapour (wavelengths are given in angstroms).



**Fig. 323.**  
A small region of the iron spectrum (from 4143 to 4236 Å).



**Fig. 324.**  
Spectrum of iodine vapour.

### 20.3. Origin of Different Types of Spectra

Investigations showed that the type of spectrum is determined by the nature of a luminous object.

*Continuous spectra* are obtained from glowing solid or liquid bodies. Incandescent particles of carbon glow in the flame of a candle, while a red-hot metal filament glows in an incandescent lamp. The same spectra are obtained with molten metals as well as glowing vapours or gases having a high density, i.e. under a high pressure. In particular, the continuous spectrum of the Sun is apparently produced by glowing high-density vapours.

*Line and band spectra* are typical of low-density glowing vapours and gases. Line spectra are emitted by glowing atoms. Many gases consist of individual atoms like the vapours of metals and inert gases (helium, neon, argon, etc.). Gases consisting of molecules, like hydrogen, oxygen, iodine vapour, may dissociate into atoms as a result of excitation. Such atomic gases produce line spectra. However, the luminescence of molecules as a whole, i.e. undissociated molecules, can also be observed. In such a case, band spectra are emitted. An excitation of such polyatomic gases or vapours sometimes leads to a partial dissociation resulting in a combination of a line and a band spectrum.

The glow of atoms and molecules in vapours and gases can be caused by heating. For example, in the flame of a gas burner one may observe the bands corresponding to the glow of a cyanogen molecule (CN), which is formed by carbon and nitrogen atoms. If we introduce a grain of common salt (sodium chloride, NaCl) into the flame, the flame acquires an intense yellow colour. A spectroscopic instrument reveals two closely spaced lines in the yellow region of the spectrum, which are typical of sodium vapour spectrum. This means that the molecules of sodium chloride in the flame of the burner have dissociated into the sodium and chlorine atoms. The glow of sodium atoms can easily be observed, while that of chlorine atoms is difficult to excite and is usually very weak. More often, atomic and molecular spectra are excited by means of electric discharges in gases. In this case, the tube with electrodes through which an electric current is passed is filled with a gas at a low pressure.

Under these conditions, a glow discharge is observed (see Vol. 2, Sec. 8.10). Sometimes, the tube for observing a glow discharge is given the shape shown in Fig. 325 in order to concentrate the glow in a narrow region, which is convenient for illuminating the slit of a spectrograph. In this figure, 1 are the electrodes, while 2 is the narrow region where the current density (i.e. the current per unit cross-sectional area) and the brightness of the glow have the maximum values. An electric spark or an arc between two electrodes under investigation may serve the same purpose.

If we increase the pressure of a glowing vapour or gas, the spectral lines

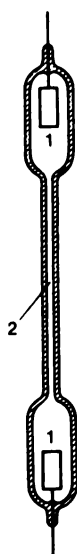


Fig. 325.  
Glow-discharge tube.

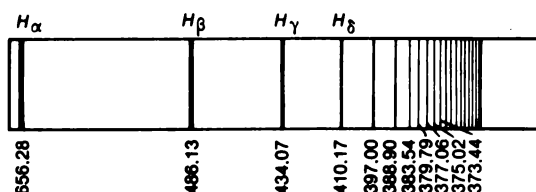


Fig. 326.  
Line spectrum of hydrogen (the Balmer series, wavelengths are given in nanometres).  $H_\alpha$ ,  $H_\beta$ ,  $H_\gamma$  and  $H_\delta$  denote the first four lines of the series, lying in the visible region of the spectrum.

will broaden and cover larger and larger spectral interval. For very high pressures (of hundred atmospheres and more), the line spectrum gradually becomes a continuous spectrum typical of compressed gases.

#### 20.4. Spectral Laws

The line spectrum of an atom is a set of a large number of lines distributed over the spectrum without any apparent order. However, a thorough analysis of spectra revealed that the arrangement of lines follows definite regularities. Naturally, these regularities are best pronounced in comparatively simple spectra typical of simple atoms. The first law of this type was established for the hydrogen spectrum depicted in Fig. 326.

In 1885, the Swiss physicist and mathematician Johann Jakob Balmer (1825-1898) established that the frequencies of individual lines in the hydrogen spectrum are expressed by the simple formula

$$\nu = R \left( \frac{1}{2^2} - \frac{1}{m^2} \right),$$

where  $\nu$  stands for the frequency of light, i.e. the number of waves emitted per unit time,  $R$  is the *Rydberg constant* equal to  $3.28984 \times 10^{15} \text{ s}^{-1}$  and  $m$  is an integer. If  $m$  is given the values of 3, 4, 5, ..., we obtain the values coinciding with the frequencies of consecutive lines in the hydrogen spectrum. The set of these lines constitutes the *Balmer series*.

Later, it was found that the hydrogen spectrum also contains a large number of lines forming series similar to the Balmer series. The frequencies

of these lines can be expressed by the formulas

$$\nu = R \left( \frac{1}{1^2} - \frac{1}{m^2} \right), \quad \text{where } m = 2, 3, 4, \dots \text{ (Lyman series),}$$

$$\nu = R \left( \frac{1}{3^2} - \frac{1}{m^2} \right), \quad \text{where } m = 4, 5, 6, \dots \text{ (Paschen series),}$$

where  $R$  has the same numerical value as in the Balmer formula. Thus, all hydrogen series can be combined in a single formula

$$\nu = R \left( \frac{1}{n^2} - \frac{1}{m^2} \right),$$

where  $n$  and  $m$  are integers such that  $m \geq n + 1$ .<sup>1</sup>

The spectra of other atoms are much more complicated, and the distribution of lines in these spectral series is not as simple as for hydrogen. It turned out, however, that the spectral lines of all atoms can be combined into series. It is very important that serial regularities for all atoms can be represented in the form similar to the Balmer formula, the constant  $R$  having nearly the same value for all the atoms.

The existence of spectral laws common for all atoms undoubtedly indicated a deep relationship between these regularities and the basic features of atomic structure. Indeed, the Danish physicist Niels Bohr (1885-1962) found in 1913 a key for understanding these regularities and at the same time established the fundamentals of the modern theory of atom (see Chap. 22).

## 20.5. Spectral Analysis Using Emission Spectra

Every atom emits its own spectral lines constituting a spectrum. Different atoms sometimes have individual accidentally coinciding lines, but the spectrum of an atom as a whole is just typical of this atom. For this reason, the emergence of a set of spectral lines belonging to an atom is a reliable indication of the fact that the given element is present in glowing vapours of a source. This important rule was formulated for the first time by the German physicist Gustave-Robert Kirchhoff (1824-1887) and the German chemist Robert Wilhelm Bunsen (1811-1899) in 1859 and formed the basis for developing the spectral method of chemical analysis. Using this method, it is possible to detect an element we are interested in even when the amount of this element is very small. An admixture of a substance whose mass is just  $10^{-7}$ - $10^{-8}$  g can reliably be established. In some especially favourable

---

<sup>1</sup> In addition to the above three series, three other series were observed in the hydrogen spectrum: the Brackett series ( $n = 4$ ), the Pfund series ( $n = 5$ ) and the Humphreys series ( $n = 6$ ).

ble conditions, substances with a mass below  $10^{-10}$  g can be detected as well.

By using this method, Kirchhoff and Bunsen made a number of important discoveries. Analyzing the spectrum of a mixture of alkali metal compounds (of lithium, sodium and potassium), they discovered that the spectrum contains a number of new lines in addition to those belonging to the known metals. They proposed that new unknown chemical elements were present in the mixture. Indeed, by appropriate treatment two new elements called *rubidium* and *cesium* were separated. Later, a few more unknown elements were discovered by using the spectral analysis (*thallium*, *indium* and *gallium*).

It is interesting to note that the element called gallium was predicted by the Russian chemist D. I. Mendeleev (1834-1907) as eka-aluminum. Mendeleev described the proposed properties of this element and indicated that it should be sought for with the help of spectral analysis. Below, we compare Mendeleev's predictions with the description given by the French chemist Paul Émile Lecoq de Boisbaudran (1838-1912) who discovered and investigated this new element.<sup>2</sup>

Properties of eka-aluminum (Ea) predicted by Mendeleev

1. Atomic weight is about 68.
2. It is a *metal* with a specific gravity of 5.9 and a low melting point. It is not volatile. Is not oxidized in air. Will decompose water vapour at red-hot temperature. Will dissolve in acids and alkalies.
3. Its *oxide* has a formula  $Ea_2O_3$  and a specific gravity of 5.5. Will dissolve in acids, forming the salt of the type  $EaX_3$ . Hydroxide will dissolve in acids and alkalies.
4. Its *salts* will have a tendency to form basic salts. Sulphates will form alums. Sulphides will be precipitated under the action of  $N_2S$  or  $(NH_4)_2S$ . Anhydrous chloride will be more volatile than zinc chloride.
5. In all probability, the element will be discovered with the help of spectral analysis.

Properties of gallium (Ga) described by de Boisbaudran

1. Atomic weight is 69.9.<sup>3</sup>
2. It is a *metal* with a specific gravity of 5.94 and a melting point of  $30.15^\circ$ . It is not volatile at moderate temperature. Does not change in air. Action on water vapour is unknown. Weakly soluble in acids and alkalies.
3. Its *oxide* is  $Ga_2O_3$  and its specific gravity is unknown. Dissolves in acids, forms the salts of the type  $GaX_3$ . Hydroxide dissolves in acids and alkalies.
4. Its *salts* are easily hydrolyzed to produce basic salts. Alums are unknown. Sulphide is precipitated under the action of  $N_2S$  and  $(NH_4)_2S$  under special conditions. Anhydrous chloride is more volatile than zinc chloride.
5. Gallium was discovered with the help of a spectrocope.

<sup>2</sup> We retained the terminology of original articles. At present, the relative atomic mass  $A_r$ , instead of atomic weight and density  $\rho$  instead of specific gravity are used. — *Eds.*

<sup>3</sup> According to modern data,  $A_r = 69.72$  and  $\rho = 5.904$  g/cm<sup>3</sup>.

In 1895, new spectral lines were discovered in the solar spectrum, which were ascribed to a new gas called *helium*.<sup>4</sup>

Somewhat later, a gas whose spectrum was found to be identical to the spectrum of hypothetical helium was discovered in pure form on the Earth. Thus, the hypothesis about a new element in the composition of the Sun was confirmed.

The example with helium is not only instructive, but also shows the importance of spectral analysis in determining the composition of celestial bodies inaccessible to a direct chemical investigation. Thanks to spectral analysis, we have at present a fairly complete information about the composition of the Universe. It has been established that it is constructed of the same elements which are available on the Earth. The data obtained on spacecrafts and artificial satellites confirm and complement the information about the composition of the Moon and other planets.

The presence of some spectral lines in a spectrum is a reliable evidence of the presence of a certain element in the mixture under investigation. This forms the basis of *qualitative analysis*. On the other hand, the measurement of intensity of a spectral line makes it possible to judge about the amount of a given element in a test sample. This problem is much more complicated since although the intensity of spectral lines increases with the concentration of a given element, the relationship between the line intensity and concentration is rather difficult. There are many factors which may affect the line intensity at a constant concentration. For this reason, the methods of analysis allowing us to determine the concentration of an element we are interested in by using spectral analysis, i.e. to carry out *quantitative analysis*, have been developed only recently.

These methods are of great practical importance since they allow us to rapidly analyze the composition of complex alloys widely used in modern engineering. Many alloys (like different grades of steel) do not differ in appearance, and their composition can be determined only by analyzing their spectra. Since the use of unsuitable grade of steel for manufacturing an important part of a machine may lead to a reject or even accident, an error in the choice of materials is extremely dangerous. For this reason, before using a steel in an industrial process, it is usually subjected to a rapid spectral analysis which takes about 1 min. The method of spectral express analysis was developed by a group of Soviet scientists headed by Academician G. S. Landsberg (1890-1957) at the end of the thirties.

Similarly, spectral analysis is used for determining the composition of ores and minerals, which makes it possible to accelerate and simplify the exploration of mineral resources and solve a number of other practical problems.

---

<sup>4</sup> Helium is derived from the Greek *hēlios* meaning the Sun + *ius*.

### 20.6. Absorption Spectra of Liquids and Solids

If the light from an incandescent lamp passes through a coloured glass or a dye solution, its colour changes. An analysis of the spectrum of such a filtered light shows that some spectral regions corresponding to the wavelengths absorbed by the coloured substance are absent or weakened. Such a spectrum is known as the *absorption spectrum*.

The form of an absorption spectrum depends on the absorbing substance. For different substances, absorption regions are formed in different parts of the spectrum and have different intensities and widths. In many cases, from the form of an absorption spectrum of a solution it is possible to determine the absorbing substance causing this spectrum, i.e. to analyze the solution. In most cases, however, the absorption spectra of solid and liquid bodies or solutions have the form of wide bands embracing the larger part of the spectrum and overlapping to a considerable extent. For this reason, it is sometimes difficult to distinguish between two absorbing substances by using their absorption spectra. Nevertheless, the practical methods of analysis of absorption spectra play more and more important role. An analysis of the ultraviolet and infrared regions of spectrum besides the visible region ensures more reliable results in this field.

### 20.7. Absorption Spectra of Atoms.

#### Fraunhofer Lines

The spectra of metal vapours consisting of individual atoms are the most typical absorption spectra. Let us direct the light from an incandescent lamp through a vessel with sodium vapour. It will be seen that the continuous spectrum of the lamp will contain two narrow black lines just in the region where the two narrow emission lines of the glowing sodium vapour are located (Fig. 327). This observation was made by Kirchhoff who established the general law according to which *the lines of absorption of atoms exactly correspond to their emission lines*. Thus, the absorption spectra of atoms are as typical of them as the emission spectra and can be used for qualitative analysis.

Kirchhoff's law made it possible to explain an important observation made by the German physicist Joseph von Fraunhofer (1787-1826) who examined in 1817 the solar spectrum with the help of a spectroscope with a

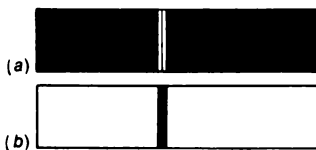


Fig. 327.

Emission spectrum (a) and absorption spectrum (b) of sodium vapour (schematic diagram).



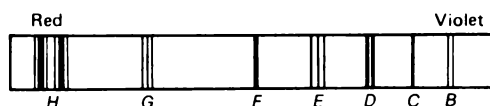


Fig. 328.

Solar spectrum with the Fraunhofer absorption lines.

diffraction grating, made by him. He paid attention to the fact that the continuous solar spectrum contains a considerable number of black lines. Fraunhofer established that these lines are not accidental and are always present in the solar spectrum in strictly definite regions (Fig. 328). These lines were called the *Fraunhofer lines* and could not be explained satisfactorily up to the discovery of Kirchhoff's law. According to this law, the Fraunhofer lines are just absorption lines of vapours of different metals arranged between the source of continuous spectrum (the bright surface of the Sun called the *photosphere*) and the spectroscopic instrumentation. These vapours constitute the solar atmosphere which is less dense and less hot than the photosphere. Thus, the solar spectrum provides information about the absorption spectra of these vapours.

Using Kirchhoff's law and comparing the position of the Fraunhofer lines with emission lines of different elements, it is possible to determine the elements constituting absorbing vapours. Thus, the composition of the solar atmosphere has been established, and hence the presence of a number of elements in the Sun.

It should be noted that the spectral analysis based on the absorption spectra of vapours is of no less importance for astronomy than the analysis based on emission spectra since it allows one to analyze the composition of bodies emitting a continuous spectrum and surrounded by the atmosphere of vapours of elements.

## 20.8. Investigation of Red-Hot Bodies.

### Blackbody

Kirchhoff's law formulated in the previous section is a special case of a more general Kirchhoff's law according to which *the emissivity of heated bodies is proportional to their absorptivity at the same temperature*. For example, having heated metal plates painted in black and white to the same temperature, we shall see that the black plate emits more heat per square centimetre than the white plate. This experiment can conveniently be made if we pour hot water into a tin cube some of whose faces are painted black and the others white. The difference in radiation can be established by bringing a hand or cheek close to these faces or using a more sensitive heat receiver like a gas thermometer (see Vol. 1, Sec. 13.15). Having made the bulb of this thermometer in the form of a flat box one of whose surfaces

is painted black and the other white, we can see with the help of this instrument that the black surface absorbs radiation better than the white surface.

In the above experiments, Kirchhoff's law is verified for the total radiation comprising various wavelengths. Finer experiments show that this law is also valid for narrow spectral regions. According to experiments, a red-hot body emits only the wavelengths which it can absorb *at the same temperature*.

A simple experiment with a gas burner may illustrate this law qualitatively. A "colourless" flame of the gas burner has no colour and glows weakly since the substances heated to high temperatures in this flame (water vapour, carbon monoxide CO and carbon dioxide CO<sub>2</sub>) absorb very weakly and hence emit visible light weakly. If, however, we introduce a grain of common salt into the flame (sodium chloride NaCl), the flame immediately becomes bright-yellow since it now contains heated sodium vapour which is a good absorber and hence a good emitter of the waves corresponding to the yellow colour. Introducing other elements into the colourless flame of the burner, we can observe how the flame acquires different colours in accordance with Kirchhoff's law. Having reduced the air supply to the burner, we obtain a bright flame since carbon, being one of the components of town gas, cannot be oxidized completely and remains in the form of thin particles of coal which is a good absorber and hence a good emitter of various wavelengths which produce white light upon mixing.

The stipulation that absorptivity and emissivity should refer *to the same temperature* is very important since the ability of a substance to absorb and emit may strongly depend on temperature. For example, a rod of molten quartz is completely colourless, and hence does not absorb visible rays. It may seem that, on the basis of Kirchhoff's law, this rod cannot emit visible light irrespective of its temperature. Experiments show, however, that at a temperature of about 1500 °C, the rod made of molten quartz glows brightly like a red-hot platinum wire. The reason behind this is, naturally, not in the violation of Kirchhoff's law but in the fact that at about 1500 °C molten quartz absorbs visible light nearly as well as a metal. In other words, it is practically opaque for visible rays at this temperature, while it is completely transparent at room temperature.

Since according to Kirchhoff's law the emission of heated bodies is proportional to their absorptivity, the body having the maximum absorption coefficient at a given temperature will have the maximum emission. According to Secs. 8.9 and 19.8, the maximum value of the absorption coefficient is unity. In this case, a body completely absorbs the entire radiation incident on it. If the absorption coefficient is unity for all wavelengths, the body is called a *blackbody*. In any spectral region a blackbody emits more energy than any other body at the same temperature. The properties of a blackbody for a rather broad spectral region (from infrared to ultraviolet radiation) are exhibited by a surface coated by a layer of soot or, even

to a larger extent, by a cavity blackened with smoke from inside and having a small aperture (cf. Sec. 19.12).

## 20.9. Temperature Dependence of Emission of Red-Hot Bodies. Incandescent Lamps

The emission of red-hot bodies strongly depends on their temperature. Having connected an incandescent lamp in series with a rheostat in a circuit and controlling the current, we can gradually increase the temperature of the filament. It will be seen that *the brightness of the filament rapidly increases with temperature*. Besides, the change in the colour of the heated filament can be noticed: from dark-red it gradually becomes bright-white. Hence it follows that the fraction of short waves rapidly increases in the radiation of the filament with increasing temperature.

A detailed analysis shows that the larger portion of energy emitted by the incandescent lamp corresponds to invisible infrared rays. As the temperature rises, the total radiant energy considerably increases, but the intensity of visible rays increases more rapidly so that their contribution to the total radiation soon becomes significant. For example, as the temperature of a platinum filament increases from 1000 °C to 1100 °C, the total radiant energy increases by a factor of 1.5, while the energy corresponding to green rays increases twenty-fold. Thus it becomes clear that an increase in the filament temperature is very expedient when the incandescent lamp is used as a light source since in this case the energy emitted in the form of visible light increases much more rapidly than the total radiant energy.

The ratio of the energy corresponding to visible rays to the entire energy spent for heating is called the *efficiency*, or *economic efficiency of a lamp*. The following table illustrates the dependence of the efficiency  $k$  on the thermodynamic temperature  $T$  of a blackbody.

| $T, K$  | 2000 | 2250 | 2500 | 2750 | 3000 | 3500    |
|---------|------|------|------|------|------|---------|
| $k, \%$ | 0.4  | 0.85 | 1.6  | 2.4  | 3.5  | about 5 |

It follows from the table that the efficiency of a lamp is generally low, but rapidly increases with temperature.

A great progress in manufacturing incandescent lamps<sup>5</sup> was associated with a changeover from carbon filaments that could not be heated to above

---

<sup>5</sup> Lamps with a metal filament (made of tungsten, molybdenum, etc.) were proposed and patented in 1890 by A. N. Ladygin (1847-1923) who invented the electric incandescent lamp in 1873 (see Vol. 2, Sec. 4.7).

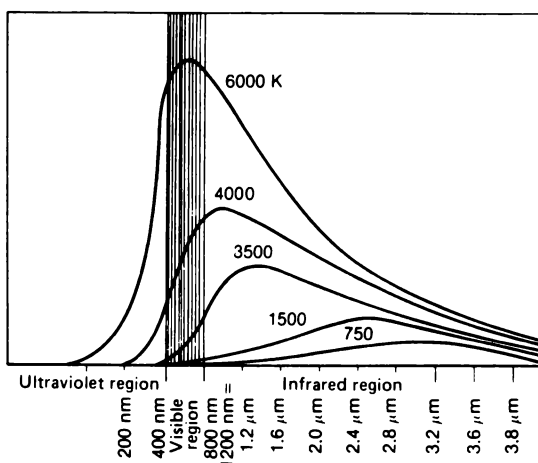
2100 K to the lamps with tungsten filaments that can be heated without fast breakdown to 2500 K. The filling of the bulbs of incandescent lamps with nitrogen or argon prevents from a rapid atomization of the filament and makes it possible to use tungsten filaments at a temperature of about 3000 K (gas lamps).

When a broken filament is accidentally brazed together as a result of a shock, a sharp increase in the brightness and a whiter colour of the lamp can be noticed. The reason behind this is that the filament becomes shorter, its resistance drops, the current through the lamp increases and the temperature and efficiency rise. The lamp becomes more economical, but in a very short time after such an abnormal operating conditions the filament breaks down. At present, the lamps filled with iodine vapour along with nitrogen and krypton are being used. This noticeably reduces the sublimation of the metal of red-hot filament and makes it possible to raise the filament temperature.

## 20.10. Optical Pyrometry

Continuous emission spectra of incandescent bodies differ but little from one another, and hence are not suitable for analyzing the structure of bodies. However, the investigation of the energy distribution in the spectrum of a red-hot body leads to important conclusions. This energy distribution is similar for different bodies. We shall limit ourselves to an analysis of radiation emitted by incandescent coal.

Figure 329 gives an idea about the energy distribution in the carbon spectrum and about its temperature dependence. The curves show that the radiation embraces not only the visible



**Fig. 329.**

Energy distribution curve in the emission spectrum of the blackbody for different temperatures: the radiation intensity is plotted along the ordinate axis and the wavelength, along the abscissa axis.

part of the spectrum but also the infrared and ultraviolet regions, the maximum of radiant energy corresponding to infrared rays for the most of temperatures represented in the figure. The region corresponding to visible rays is hatched. It can be seen from the figure that this region constitutes a small fraction of the entire radiation. As the temperature rises, the total radiant energy increases (the curves become higher) and in accord with Sec. 20.9, the fraction of visible radiation noticeably increases.

It is noteworthy that as the temperature rises, the region corresponding to the maximum radiation is shifted to shorter wavelengths. A thorough analysis of this phenomenon reveals that the position of the maximum depends only on the temperature of an emitting body.

Strictly speaking, these conclusions are applicable only to the blackbody radiation. However, they can *successfully* be used for the radiation of red-hot metals and the solar radiation. This allows us to use the above law for solving an important problem of determining the temperature of glowing bodies. The application of this law to the Sun shows that the maximum of the solar radiation lies at about 500 nm, i.e. in the yellow-green region of the spectrum, corresponding to a temperature about 5800 K. This is the *effective temperature* of the Sun which characterizes its surface and, obviously, gives no information about its interior where the temperature apparently reaches millions of kelvins.

This method of determining the temperature of red-hot bodies is used for solving both scientific and technical problems and is known as *optical pyrometry*. It helps determine the temperature of the red-hot filament of incandescent lamps, the temperature of molten metals in blast furnaces, and so on.

## Chapter 21

# Effects of Light

### 21.1. Action of Light on a Substance.

#### Photoelectric Effect

A light wave incident on a body is partially reflected from it, partially passes through it and is partially absorbed from it (see Sec. 8.9). In most cases, the energy of absorbed light wave is transformed into the internal energy of a substance, which results in heating of the body. Sometimes, however, some portion of this absorbed energy may cause other phenomena as well. Very important effects of light which have found wide practical applications are *photoelectric effect*, *photoluminescence* and *photochemical transformations*.

A simple experiment revealing the *photoelectric effect* was described in Vol. 2, Sec. 1.9. A thoroughly cleaned zinc plate *1* (Fig. 330) is fixed to electroscope *2* and illuminated by source *3* rich in ultraviolet radiation (electric arc or a quartz mercury-vapour lamp). If the electroscope has a negative charge, under the action of light from the mercury lamp it loses the charge. The discharging occurs the more rapidly, the higher the illuminance of the plate, i.e. the larger the luminous flux incident on it. No discharging is observed if glass *4* absorbing ultraviolet radiation is arranged on the way of the rays. If the electroscope bears a positive charge, it remains on it in spite of the illumination.

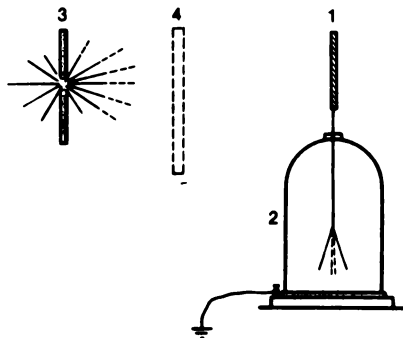


Fig. 330.

Photoelectric effect: a metal loses negative charges under the action of light.

These experiments, as well as other similar ones, lead to the following conclusions. *A negative charge disappears from the surface of a metal upon illumination. A positive charge remains on the surface of the metal in spite of illumination.* This important conclusion indicating that the effect is observed only when the illuminated plate is connected to the negative pole of a battery was definitely established for the first time by the Russian physicist A. G. Stoletov (1839-1896). For a zinc plate, the illumination of ultraviolet rays is of major importance.

The experiment described above shows the difference in the properties of negative and positive charges constituting a metal. The former are electrons weakly bound to the metal and able to easily move in the metal (conductivity) and readily move beyond its limits (photoelectric effect). The latter are positive ions constituting the crystal lattice of the metal so that their knocking out is just the sublimation of the metal. If the metal is negatively charged, a released electron is removed from the metal under the action of the electric field produced by the charged metal. In the case of a positive charge, the electrons that are always present in the metal could also be freed by light. But the electric field produced around the positively charged body decelerates escaping electrons and tends to return them back to the body. Therefore, if the kinetic energy of an escaping electron (and hence its velocity) is not too high, despite the effect of light the electrons cannot leave the plate, and its positive charge remains unchanged.

The ability of light to cause the separation of electrons from a metal is one of the best proofs of the electromagnetic nature of light waves. Under the action of the electric field of a light wave, an electron acquires an energy sufficient for leaving the limits of the metal in spite of the action of forces keeping it. However, an analysis of the laws of photoelectric effect shows that the situation is much more complicated.

## 21.2. Laws of Photoelectric Effect

As follows from Sec. 21.1, the photoelectric effect is characterized by the number of electrons released by light per unit time (i.e. by the photoelectric current) and by the velocity of these electrons. The larger the number of electrons escaping per unit time, the shorter the time during which the electrometer is discharged. The higher the velocity of electrons, the stronger the decelerating field that must be applied to prevent their escaping from the plate. The experiment shown schematically in Fig. 331 can be used for measuring these two important characteristics of the photoelectric effect, viz. the current and electron velocity.

Plate 1 from which photoelectrons are released is connected to a battery pole, the other pole being connected through a potentiometer and galvanometer to plate 2. The two plates 1 and 2 are contained in a vessel

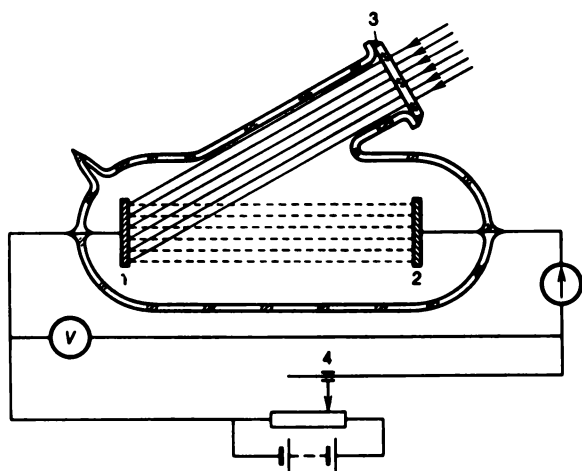


Fig. 331.

Schematic diagram of the experiment on measuring the photoelectric current and the velocity of photoelectrons: 1 — illuminated plate (cathode), 2 — auxiliary electrode (anode), 3 — window transparent for ultraviolet radiation, 4 — slider of the potentiometer.

from which air has been pumped out in order to eliminate the complications introduced into observed phenomena by collisions between electrons and gas molecules and to protect the plates from oxidation. The ultraviolet radiation incident on plate 1 penetrates through a quartz window 3. The electrons escaping from plate 1 get into the electric field applied between the plates. The voltage across the plates can be varied by moving slider 4 of the potentiometer.

If the field is strong enough and directed so that it entrains electrons from plate 1 to plate 2, all the electrons leaving plate 1 reach plate 2, and hence the current through the galvanometer is determined by the number of electrons liberated by light per unit time. This current, known as the *saturation current*, just determines the photoelectric current. If, however, the field decelerates the electrons, by increasing its magnitude we can stop all the electrons escaping from the plate. The strength of the cutoff field can be used for determining the velocity of escaping electrons.

Let us suppose that the velocity of an escaping electron is  $v$ , its mass is  $m$  and the charge is  $-e$ .<sup>1</sup> The kinetic energy of such an electron is  $mv^2/2$ . The electron having this energy can fly through a decelerating field produced by the potential difference  $U$  if  $eU$  is less than or equal to  $mv^2/2$ . Having determined this minimum value of  $U$  that retards the electrons

<sup>1</sup> The letter  $e$  here stands for the elementary charge, i.e. the positive charge equal to the electron charge in magnitude. The electron charge is negative and equal to  $-e$ . — Eds.



liberated by light, we can find the velocity of these electrons from the condition

$$\frac{1}{2}mv^2 = eU, \quad v = \sqrt{\frac{2eU}{m}}.$$

Experimental investigations similar to that described above lead to the following laws of photoelectric effect.

1. *The number of electrons freed by light per unit time (i.e. the saturation current) is directly proportional to the luminous flux.*
2. *The velocity of escaping photoelectrons does not depend on the illuminance but is determined by the frequency of light.*

The experiment shown schematically in Fig. 331 cannot be used for accurate measurements. If the distance between the plates is large in comparison with their dimensions, it is impossible to capture all the electrons released by light (i.e. to obtain the actual value of the saturation current) and difficult to establish the exact value of  $U$ , which determines the velocity of photoelectrons. A more perfect instrument was proposed by P. I. Lukirskii. Here the electrodes form a *spherical capacitor*, one electrode being made in the form of a small ball at the centre of a sphere whose surface forms the other electrode. Such an instrument makes it possible to reliably determine the saturation current and the retarding potential  $U$ , and hence to determine the photoelectric current and the maximum velocity of escaping electrons.

A natural question arises as to how the number of electrons freed by light and their velocity depend on the material of an illuminated body.

An analysis of the escape of electrons from heated metals (see Vol. 2, Sec. 7.4 and 7.5) showed that every substance is characterized by its *work function*, i.e. each metal is characterized by a certain energy which has to be imparted to an electron to enable it to overcome the forces keeping it within the metal. We arrive at exactly the same conclusions by analyzing the emission of electrons under the action of light. For some metals, the work function could be determined both by analyzing the emission of electrons as a result of heating and as a result of photoelectric effect. The two methods yielded the same results. For example, the following values were obtained for the work function of tungsten:

$$\begin{aligned} 7.18 \times 10^{-19} \text{ J} & \text{ for photoelectric emission,} \\ 7.23 \times 10^{-19} \text{ J} & \text{ for thermionic emission.} \end{aligned}$$

Let us suppose that the light of frequency  $\nu$  liberated from a metal having a work function of  $A$  electrons having a velocity  $v$ , i.e. a kinetic energy of  $mv^2/2$ . Thus, the total energy imparted to each electron is  $W = A + mv^2/2$ . The experiments similar to those described above showed that the total energy imparted to an electron by light is directly proportional

to the light frequency  $\nu$ , i.e.  $W = A + mv^2/2 = h\nu$ , where  $h$  is a constant. This constant is not only independent of the frequency of light and illuminance, but also has the same value for all substances. Therefore,  $h$  is a *fundamental constant* known as *Planck's constant* after the German physicist Max Planck (1858-1947). The value of  $h$  can be determined from the above experiments since the values of  $A$ ,  $mv^2$  and  $\nu$  can be measured. It was found that  $h \approx 6.6 \times 10^{-34}$  J·s.

Using the obtained relations, we can formulate the law of photoelectric effect as follows: *the total energy received by an electron from light of frequency  $\nu$  is equal to  $h\nu$ .*

A metal emitting electrons under the action of light must acquire a positive charge. As a result, the emerging electric field hampers the subsequent emission of electrons. What is the limiting potential difference  $U$  between the illuminated plate and the laboratory walls (ground) whose emergence prevents the subsequent escape of electrons from the plate? In the conditions of the experiment shown in Fig. 330, this potential difference is determined by the readings of the electrometer. The above question can be answered by using the basic relations given above:

$$A + \frac{1}{2}mv^2 = h\nu, \quad eU = \frac{1}{2}mv^2,$$

where  $e$  is the elementary charge.<sup>2</sup> Performing appropriate calculations for the tungsten plate (for which, as was mentioned above,  $A \approx 7.2 \times 10^{-19}$  J) illuminated by the ultraviolet radiation with a wavelength of  $\lambda = 200$  nm, we obtain  $U \approx 1.7$  V. In other words, in order to observe in experiment the metal acquiring a positive charge under the effect of radiation, we must have a sensitive electrometer or operate with a very short-wave radiation, say, X-rays (see Ex. III.37). Having found experimentally the value of  $U$ , we can use these data for determining the wavelength of the X-rays.

### 21.3. Light Quanta

The law formulated at the end of the previous section brings about quite new features to the concept of light. It states that the light of frequency  $\nu$  imparts an energy of  $h\nu$  to an electron irrespective of the luminous intensity. For a more intense light, a larger number of electrons receives the indicated portions of energy, and a smaller number of electrons receives these portions from a weak light, the portions themselves remaining equal to  $h\nu$ .

Thus, the atomistic nature is ascribed to the luminous energy. The energy of light of a given frequency  $\nu$  cannot be divided into arbitrary parts but is manifested in the form of strictly definite equal portions, i.e. "atoms" of luminous energy. A special name was given to these portions of energy.

---

<sup>2</sup> The relation  $A + mv^2/2 = h\nu$  is known as Einstein's formula. It is valid for the maximum velocity of electrons escaping from a plate under the effect of radiation of frequency  $\nu$ . Due to different factors, not all electrons leaving the plate have this maximum velocity.

They are called *light quanta*, or *photons*. The concept of light quanta was introduced in 1905 by Einstein.<sup>3</sup>

The fact that most of optical experiments do not reveal the quantum nature of luminous energy is not surprising. Indeed,  $h$  has a very small value approximately equal to  $6.6 \times 10^{-34}$  J·s. Let us calculate the energy of a quantum of green light, say, for  $\lambda = 500$  nm. The corresponding frequency  $\nu = c/\lambda = 3 \times 10^8 / 5 \times 10^{-7} = 6 \times 10^{14}$  Hz. Consequently,  $h\nu = 4 \times 10^{-19}$  J, which is a very small quantity. The energy we deal with in the majority of experiments consists of a very large number of quanta. Naturally, it is impossible to note that the energy is always equal to an integral number of quanta. Similarly, most of experiments with conventional portions of a substance always involve a very large number of atoms of the substance. For this reason, we cannot notice in these experiments that a given substance consists of an integral number of the minimum portions, viz. atoms. Special experiments are required to reveal the atomic structure of matter clearly. In the same way, in most of conventional optical experiments the circumstance that the luminous energy consists of individual light quanta remains unnoticed. However, in special experiments including the above-described experiments on photoelectric effect, *the quantum nature of light is manifested quite clearly*.

Not only photoelectric effect, but also many other phenomena in optics, atomic and molecular physics are of quantum nature.

It is extremely important that in all these phenomena we deal with the fundamental constant denoted by  $h$ . This constant is now determined from measurements pertaining to quite diverse phenomena. The numerical values determined in this way are in excellent agreement with one another.

The concept of light quanta provides an easy explanation to the basic law of photoelectric effect, viz. *the proportionality of the luminous flux and the photoelectric current*. The luminous flux, i.e. the energy transferred by light per unit time, is determined by the number of light quanta arriving in a unit of time. Obviously, the larger this number, the larger the number of electrons acquiring the additional energy carried by these quanta, and the larger the number of electrons escaping from an illuminated metal per unit time, i.e. the stronger the photoelectric current. Naturally, this does not mean that the number of escaping electrons must be equal to the number of quanta getting into the metal during this time. Not every quantum imparts its energy to an individual electron. A considerable portion of energy is distributed among the atoms of the metal, which leads to its heating.

<sup>3</sup> Albert Einstein (1879-1955) was an outstanding scientist, one of the founders of modern physics. He was born in Germany and also worked in Switzerland. When Nazis came to power in Germany, he emigrated to the USA.

Indeed, experiments show that only a small fraction (about 1%)<sup>4</sup> of the luminous energy is usually converted to the energy of escaping electrons. The remaining part of the absorbed light quanta is spent for heating the metal.

#### 21.4. Application of Photoelectric Phenomena

An analysis of the laws of photoelectric effect helped deepen our knowledge about light. For this reason, photoelectric phenomena are of great scientific importance. At the same time, the practical (technical) value of photoelectric effect is also significant. The applications of photoelectric effect have become especially diverse after photoelectric cells, or photocells, sensitive not only to ultraviolet radiation (as was described in Sec. 21.3) but also to infrared radiation and to visible light were created.

The relation  $A + mv^2/2 = h\nu$  indicated that as  $\nu$  is reduced, i.e. the wavelength of incident light increases, the velocity of knocked-out electrons decreases. When  $\nu = A/h$ ,  $v = 0$ . This means that at the corresponding frequency, electrons cannot be liberated from the metal, and photoelectric effect does not take place.

Thus, for every metal there exists a limiting wavelength of light which is able to cause the photoelectric effect. If the incident light has a wavelength larger than the limiting value, no photoelectric effect emerges however intense is the light. For this reason, for example, we had to resort to ultraviolet radiation to observe the photoelectric effect for zinc since the work function for zinc is quite large ( $A_{Zn} = 6.8 \times 10^{-19}$  J). For other materials, the value of  $\lambda$  can be increased since the work function for them is smaller. Alkali metals are especially convenient for this purpose (sodium, potassium, rubidium, and the more so cesium:  $A_{Cs} = 3 \times 10^{-19}$  J). The work function becomes even smaller if the surfaces of these metals are coated with a special film. As a result, surfaces sensitive not only to visible but also to infrared light have been obtained.

Photocells convenient for practical applications are made in the form of an evacuated glass bulb with a layer of a sensitive metal on the inner surface. Sometimes, a certain amount of neutral gas (say, argon) is introduced into the bulb. The gas does not destroy the surface of the metal but can be ionized under the impact of escaping electrons and thus increase the observed current by adding its ions to it (see Vol. 2, Sec. 8.3). The surface of the sensitive metal serves as an electrode of the photocell (cathode). The anode is a metal ring or plate soldered into the bulb. Having applied

---

<sup>4</sup> As was mentioned in Sec. 7.1, at present up to 15% of luminous energy is used for obtaining photoelectric current. The figure given in the text refers to photoelectric phenomena observed in illuminating metals.

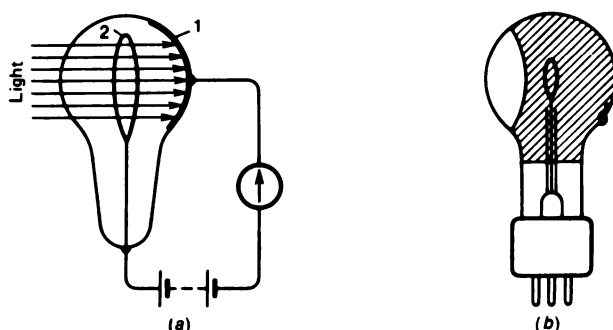


Fig. 332.

Vacuum photocell: (a) circuit diagram of connection: 1 — light-sensitive layer (cathode), 2 — anode in the form of a ring, (b) schematic diagram.

an appropriate voltage between the electrodes, we obtain a photocell ready for operation (Fig. 332).

Subsequently, the photoelectric effect emerging between a metal and an oxide film formed on its surface was also used. A thin layer of a semi-conducting material is formed between the metal and its oxide, which has the property to let through electrons escaping from the metal and prevent from their passage in the opposite direction. The explanation of the operation of this *restraining layer* is rather complicated (see Vol. 2, Sec. 9.3). The practical application of such surfaces has led to obtaining photocells having serious advantages over other types. They are more sensitive than the photocells based on the photoelectric effect from the free surface of a metal and do not require an auxiliary battery. They can be made in any shape which is the most convenient (Fig. 333). Since photoelectric current is proportional to luminous flux, photocells are widely used as photometers of various designs. One of such photometers intended for determining illuminance (luxmeter) is described in Sec. 8.11. The possibility to register light signals with the help of electrical measuring instruments allows us

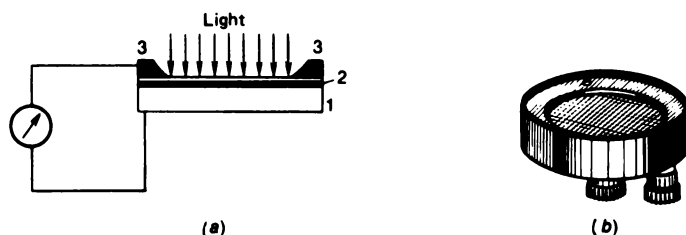


Fig. 333.

Barrier-layer cell: (a) circuit diagram of connection: 1 — metal, 2 — oxide film with a barrier layer; a thin (transparent) metal layer is deposited by evaporation, metal ring 3 pressed against the metal serves as the second electrode; (b) general view of a barrier-layer cell.

to combine photocells with relays (see Vol. 2, Sec. 18.12). As a result, photocells can automatically make various complex operations. Automatic devices for counting, detection, starting or terminating different operations have been constructed. The number of various applications of photocells is extremely large, and new and new devices of this type appear every day. It was mentioned in Introduction that new photocells (using semiconductors such as germanium and the more so silicon) are capable to convert a considerable portion of luminous energy into electric energy and are employed for utilizing the solar energy (solar batteries). Solar batteries having the surface area of the order of ten square metres supply energy to artificial Earth satellites.

### 21.5. Photoluminescence. Stokes' Shift

Some materials not only reflect a part of light incident on them, but also start to glow. Such a glow, or *luminescence*, has an important peculiarity: *luminescent radiation has a different spectral composition than the light causing luminescence*.

An example of easily observable luminescence is a bluish-milky glow of kerosene observed in daylight. A large number of solutions of dyes and other substances exhibit luminescence especially as a result of illumination from sources emitting ultraviolet light (say, electric arc or mercury-vapour lamp). The glow of this kind is known as *photoluminescence*<sup>5</sup>, which means that it emerges under the action of light.

The change in the colour of a glow in comparison with the colour of the exciting radiation can sometimes be noticed by eye. This peculiarity becomes even more noticeable if we compare the spectrum of the luminescent radiation with that of the exciting radiation. A comparison shows that the wavelength of the luminescent radiation is longer than that of the exciting radiation.

This rule stating that the luminescent radiation is characterized by a longer wavelength than the exciting radiation is known as *Stokes' shift* after the English physicist George Gabriel Stokes (1819-1903).

Any experiment involving the excitation of photoluminescence can serve as an illustration of this shift. For example, we can pass the light from a lantern through a violet glass absorbing practically all blue and longer waves (Fig. 334). If the beam of such a light is directed on a flask containing a solution of fluorescein, the illuminated liquid starts to glow with greenish-yellow colour.

Using sources of light whose radiation contains a considerable number of short waves (of the ultraviolet range), we can see that almost all bodies

<sup>5</sup> The word "photoluminescence" is quite an awkward combination of the Greek *phōt* meaning light and the Latin *lumen* meaning light + escence.

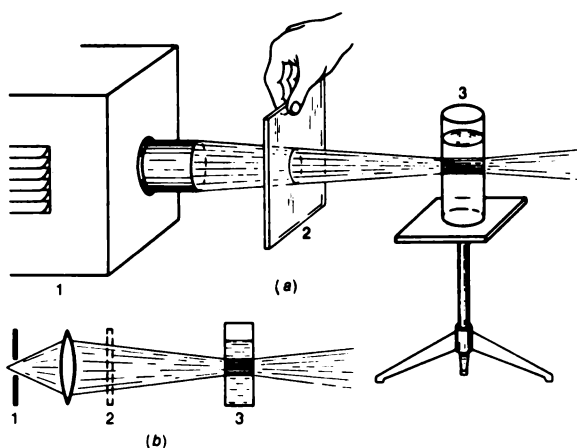


Fig. 334.

Experiments on fluorescence: (a) arrangement of devices, (b) schematic diagram of experiment. 1 — light source (lantern), 2 — light filter (violet), 3 — vessel containing a fluorescent liquid.

are able to luminescence to a certain extent. Sometimes luminescence can be strongly intensified by cooling a body to a very low temperature, say, by immersing it into liquid air.

It is noteworthy that some bodies continue to glow for some time after the illumination has ceased.

Such an *afterglow* has different durations for different materials. In some objects, it is observed for a very short time ( $10^{-4}$  s or even less) and requires special instruments for its observation. In others, it lasts for many seconds and even minutes (hours) so that it can be observed without any difficulty.

The glow which ceases with illumination is usually called *fluorescence*, while the glow of a considerable duration is known as *phosphorescence*. However, it should be borne in mind that it is difficult to distinguish between fluorescence and phosphorescence so that this division is conditional to a certain extent.

Prolonged phosphorescence is observed for many specially prepared crystalline powders. They are used for manufacturing *phosphorescent screens*. A sheet of cardboard coated, for example, with zinc sulphide becomes a good phosphorescent screen preserving its glow two or three minutes after the illumination has been switched off.

Such screens also glow under the action of X-rays. It should be noted, however, that the luminescence induced by X-rays is a more complex phenomenon than that caused by ordinary light since here the major role is played by fast electrons knocked out by X-rays.

An important modern application of phosphorescent powders is the manufacturing of fluorescent lamps. The glow induced by an electric current in gas-discharge lamps, like in a mercury-vapour tube, normally contains a large fraction of ultraviolet radiation which is not only unsuitable for illumination, but also harmful for the eye. The Soviet physicist S. I. Vavilov (1891-1951) proposed that the inner surface of such lamps be coated by specially prepared phosphorescent compound converting the ultraviolet light into visible light (in accordance with Stokes' shift). This gives a considerable economy since the amount of electric energy converted in such lamps into the energy of visible light is three times larger than in incandescent lamps. By choosing an appropriate composition of a phosphorescent substance, the spectral composition of the emitted light can also be improved to bring it closer to the spectral composition of daylight. This is employed in the construction of fluorescent lamps which are being used now on an ever widening scale.

### 21.6. Physical Meaning of Stokes' Shift

Stokes' shift can be interpreted with the help of quantum-mechanical concepts. Let us suppose that a glow is induced by a monochromatic light of frequency  $\nu$ . Thus, a molecule of a luminescent substance absorbs energy in the form of a quantum  $h\nu$ . The processes caused by the energy absorbed in the molecule are rather complex. A part of energy of the quantum is spent in these processes, the other part being emitted again in the form of luminescent light. Consequently, the emitted quantum must have a lower energy, i.e. correspond to a lower frequency  $\nu$ . This reduction in frequency (increase in wavelength) forms the basis of Stokes' shift.

The fact that the emitted light contains different wavelengths even with the excitation by a monochromatic light indicates that the processes of conversion of energy of the light quantum within a molecule are rather complex and diversified. They have not been studied completely so far, and hence the theory of photoluminescence is not quite clear.

### 21.7. Luminescent Analysis

In addition to the above-mentioned application of luminescence for phosphorescent screens and various luminous paints used for decoration and show business, another important application of this phenomenon should be mentioned. Luminescence is characterized by a very high sensitivity: it is sometimes sufficient to have  $10^{-10}$  g of a luminescent substance in a solution to detect it from the typical glow. Luminescence is used for observing insignificant traces of a substance constituting a millionth of percent in a mixture. This high sensitivity makes luminescence an important tool for detecting infinitesimal amounts of impurities, which enables us to judge about contamination or processes leading to changes in an initial substance.

For example, by using luminescence, we can observe the initial stage of rotting process in foodstuffs. The luminescent analysis is used for exploring oil deposits. If the soil extracted during drilling contains even insignificant traces of oil, they can easily be detected from fluorescence. In this



way, close oil-bearing layers can be recognized. There exist many other fields of application of luminescent analysis in engineering.

### 21.8. Photochemical Action of Light

The absorption of light may also cause some chemical processes which usually involve the decomposition of a molecule that has absorbed light, which frequently leads to a number of further chemical transformations. The chemical process that occurs under the action of light in green parts of plants is of special importance.

It is well known that the respiration of living organisms is accompanied by the oxidation of carbon contained in their bodies. Oxidation (burning) of carbon (its transformation into carbon dioxide,  $\text{CO}_2$ ) is accompanied by the liberation of energy which is required for movements of living organisms. Similarly, the main source of energy used in engineering is the process of combustion of a fuel, i.e. again the process of formation of  $\text{CO}_2$ .

The inverse process of decomposition of  $\text{CO}_2$  occurs in green parts of plants under the action of solar radiation and is known as a photochemical process. The decomposition of carbon dioxide is accompanied by further chemical transformations which ultimately lead to the formation of the basic organic compounds constituting the bodies of plants and animals. Thus, the "great circulation" of carbon in nature is due to photochemical transformations. The energy spent in the process by solar radiation is stored in the form of internal energy of the products of conversion and is the energy stock that has been used by man until recently.

An important role in the analysis of  $\text{CO}_2$  decomposition under the action of light was played by the investigations carried out by the Russian biologist K. A. Timiryazev (1843-1920) who established that this process is associated with chlorophyll contained in plants and responsible for the green colouring of their leaves. He found out that the process occurs mainly under the action of red radiation of the solar spectrum, which is most strongly absorbed by chlorophyll. The entire photochemical process, however, is very complicated, and in spite of the advances made in recent years, which clarified individual stages of the process, their sequence and interrelation still remain unclear.

Besides the photochemical process described above, which occurs on a giant scale, a number of other photochemical transformations are known. A simple example of this type is the photochemical process of *bleaching* of many dyes, which consists in their oxidation by atmospheric oxygen under the action of light. Having painted a layer of gelatin by a solution of a paint (cyanine), we can keep such a painted plate for a long time. If, however, we direct an intense light beam (from the Sun or an arc lamp) on it, the plate is bleached so rapidly that the regions illuminated by light become colourless immediately. The bleaching of canvas spread in full

blaze of the Sun is essentially a photochemical bleaching. Many photochemical processes are used at present to accelerate the synthesis of some substances. Most of such processes are intensified by the action of short-wave ultraviolet radiation.

### 21.9. The Role of Wavelength in Photochemical Processes

The role of light in photochemical processes is reduced to the imparting to a molecule of an energy high enough for splitting it into components. It follows from the concept of light quanta that the energy imparted by light to an individual molecule is very high and is the higher, the smaller the wavelength.

Indeed, since a molecule absorbs light in quanta, the energy absorbed by it is equal to  $h\nu$ . For a light having a wavelength of 480 nm, we have  $h\nu = 4.1 \times 10^{-19}$  J.

It is interesting to note that the average kinetic energy of an individual gas molecule attains this value only at a temperature of about 20 000 °C.

In other words, even the illumination by visible light may split molecules as effectively as heating by 20 000 °C. Consequently, the illumination with ultraviolet radiation of X-rays may turn out to be still more effective.

### 21.10. Photography

Photography is also based on a photochemical process. The sensitive layer of a photographic plate is made of gelatin with crystals of silver bromide distributed in it. Under the action of light, a molecule of silver bromide (AgBr) decomposes so that metallic silver is deposited in the form of tiny particles. If the amount of such silver particles per unit surface is considerable, the plate becomes dark. This can be observed by exposing to light a photographic plate half of which is wrapped in black paper. Having unwrapped the black paper after a long time, we shall clearly see the boundary between the unexposed (light) and illuminated (dark-grey) parts of the plate.

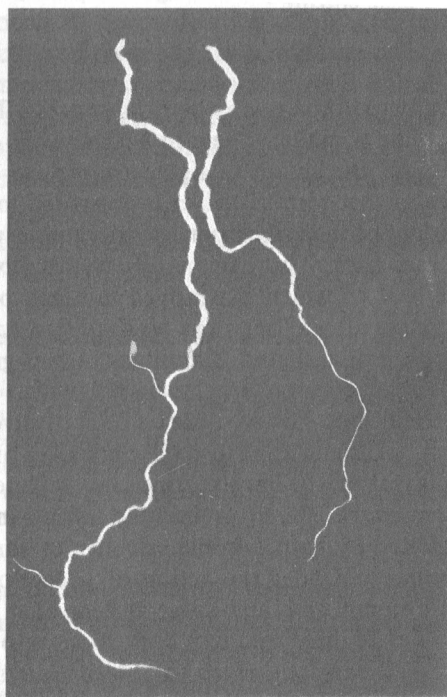
Under normal conditions, however, the amount of silver liberated under the action of light is so insignificant that the darkening of the plate cannot be observed. For this reason, the image formed in the plate but still invisible is known as the *latent image*. The silver bromide crystals in which decomposition has been initiated by light become sensitive to some chemical substances called *developers*. Under the action of a developer, a "charged" crystal of AgBr is decomposed, and silver is deposited in the form of a fine dark powder. A plate immersed in a dark room into such a developer soon becomes black in the regions which have been illuminated, the darkening being the stronger, the higher the illuminance of the corresponding region of the plate. The process of chemical treatment of an exposed plate is known as the *development*.

Having developed the plate, we must dissolve nondecomposed silver bromide in the solution of hyposulphite, and having washed the plate, we

obtain the plate with a fixed image, i.e. the plate which is no longer sensitive to the action of light.

It is clear that the plate contains a *negative image*, i.e. the illuminated regions corresponding to the bright parts of the pattern being photographed will be dark, and vice versa. Applying such a negative to a new plate or to a photographic paper which also has a light-sensitive layer, and illuminating the paper through the negative, we obtain a new image which must be developed and fixed in the same way.<sup>6</sup> This image will be *positive* since the bright parts in it will correspond to the well-illuminated regions of the pattern.

Photography plays a very important role in culture, science and engineering since it makes it possible to obtain exact images of instantaneous patterns or of objects illuminated so weakly that it is impossible to see their details by the naked eye.



**Fig. 335.**

Photograph of a lightning (obtained in Stekolnikov's laboratory).

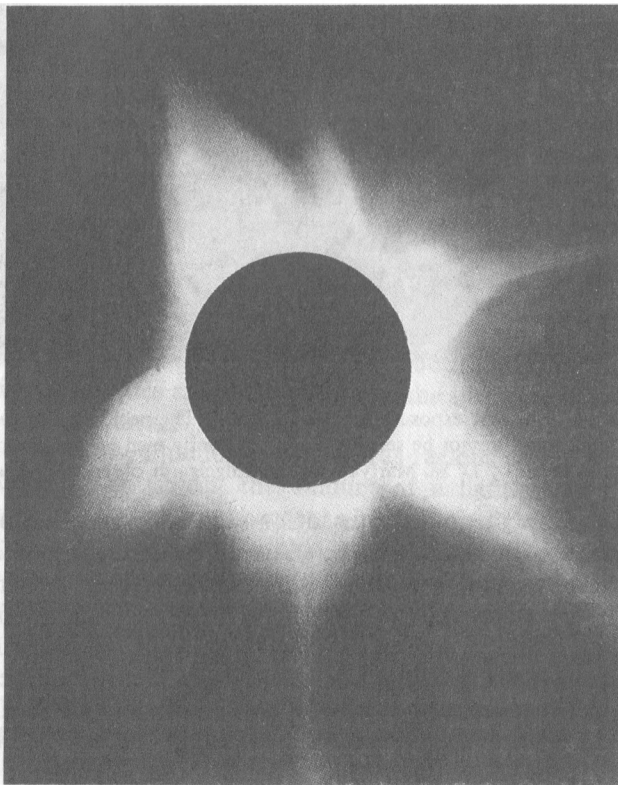
---

<sup>6</sup> In the so-called "developing-out" paper containing silver iodide, a dark image is formed in illuminated regions even without development. The image obtained on such a paper must be just fixed.

For example, it is difficult to obtain a reliable sketch of a lightning which lasts for a small fraction of a second. On the other hand, a photograph (Fig. 335) gives an idea about its exact shape. It is interesting to note that by shooting a lightning with a moving camera, one can make sure that the lightning is a series of repeating electric discharges separated by hundredths of a second so that each such discharge occurs during a thousandth of a second.

The phase of a total solar eclipse is normally very short (sometimes less than a minute). For this reason, no reliable image of the solar corona (which can be seen only during a total eclipse) could be obtained before photography had been developed (Fig. 336).

On the other hand, photography can be used for observing objects that emit a very weak luminous flux during a long time. By exposing a photographic plate to such a light for a long time, we “accumulate” the photo-



**Fig. 336.**

Photograph of the solar corona (obtained by A. A. Mikhailov during a total solar eclipse).



**Fig. 337.**

Photograph of the night sky; exposure time was 2 hours. The nebula which resembles the contours of America and cannot be seen by the eye is clearly seen on the photograph. The photograph was taken by D. Ya. Martynov at the Engelgardt observatory near Kazan.

graphic effect. This method was used for examining a region of the sky (Fig. 337). The two-hour exposure was used for obtaining the photograph. We can see the image of a nebula which cannot be seen even through a powerful telescope. Naturally, in order to retain the same orientation of the telescope relative to a celestial object in spite of the diurnal rotation of the Earth, the instrument must be rotated with the help of a clockwork mechanism in the direction opposite to that of the rotation of the Earth.

Finally, photography can be used for obtaining images of objects emitting invisible radiation (X-rays, ultraviolet or infrared radiation of up to 1200 nm). This circumstance is very important for numerous scientific investigations.

The action produced by light on a photographic plate depends on the wavelength. Simple silver bromide plates are sensitive to light starting from a wavelength of 450 nm, i.e. indigo-violet rays. Red, yellow and green rays do not affect the plate since they are not absorbed by silver bromide. Therefore, the objects of such colours are indistinguishable from black ones. This leads to a rather distorted distribution of light and dark regions on photographs.

At present, plates are also made sensitive to longer waves by applying special dyes to gelatin, which absorb these long waves and transfer the absorbed energy of silver bromide (*sensitization* of plates). This method is employed for obtaining *orthochromatic* plates which are sensitive to about  $\lambda = 600$  nm, and *panchromatic* plates sensitive to the entire visible spectrum. The plates sensitive to infrared rays up to  $\lambda = 1200$  nm are also prepared for special scientific and technical purposes.

### 21.11. Photochemical Theory of Vision

Visual sensations of human beings and animals are also associated with photochemical processes. Light incident on the retina is absorbed by light-sensitive substances (rhodopsin, or visual purple, in rods and iodopsin in cones). The mechanism of decomposition of these substances and their subsequent reduction remains unclear so far, but it was found that the products of decomposition cause a stimulation of optic nerve, as a result of which electric pulses pass through the nerve to cerebrum, and a sensation of light is produced. Since the optic nerve branches over the entire surface of the retina, the nature of a stimulation is determined by the regions of the retina where the photochemical decomposition took place. Therefore, from the stimulation of the optic nerve one can judge about the nature of the image on the retina, and hence about the pattern in the surrounding space, which is reflected in this image.

Depending on the illuminance of various regions of the retina, i.e. on the brightness of an object, the amount of a light-sensitive substance decomposed per unit time, and hence the intensity of light sensation, varies. It should be emphasized, however, that the eye is able to perceive sufficiently well the images of objects despite a drastic difference in their brightness. We clearly see objects illuminated by bright solar rays as well as the same objects in evening twilight, when their illuminance, and hence brightness (see Sec. 8.6), is lower by a factor of  $10^4$ . This ability of the eye to adjust itself to a wide range of brightness is known as *adaptation*. The adaptation to brightness is realized in several different ways. For example, the eye rapidly responds to a change in brightness by changing the diameter of the pupil, thus changing its area, and hence the illuminance of the retina, approximately by a factor of 50. The mechanism ensuring the adaptation to light

in a much wider range (by a factor of 1000) is much slower. Besides, the eye is known to have sensitive elements of two types: rods which are more sensitive to light and cones which are less sensitive to light but are able to distinguish colours. Rods play the major role in the dark (for weak illumination), ensuring the twilight vision. In bright light, visual purple in the rods is rapidly bleached, and they lose the ability to perceive light. In this case, only cones are functioning, whose sensitivity is much lower and for which new illumination conditions are quite admissible. The adaptation takes the time equal to the time of "blinding" of the rods, which is normally equal to 2-3 min. In the case of a too fast transition from darkness to a bright light, this protecting process may have no time to occur, and the eye becomes blind for a time or forever depending on the severity of blindness. A temporary loss of sight, which is well known to drivers, occurs because of glaring headlights of vehicles coming from the opposite direction.

The fact that rods and not cones operate in the case of a weak illumination (in twilight) explains why it is impossible to distinguish colours in the evening ("when candles are out all cats are grey").

As to the ability of the eye to distinguish colours in the case of a bright illumination, when the cones are put into operation, this problem cannot be assumed to be solved completely. In all probability, there are three types of cones in the eye (or three types of mechanisms in each cone), which are sensitive to three different colours, viz. red, green and blue, whose various combinations produce sensations of any colour. It should be noted that in spite of advances in this field made in recent years, the direct experiments concerning the investigation into the structure of the retina give no reliable information about the existence of the triple mechanism indicated above and forming the basis of the three-colour theory of colour vision.

The existence of two types of light-sensitive elements, viz. rods and cones, leads to another important phenomenon. The sensitivity of both rods and cones to different colours is different. But the sensitivity maximum for cones lies in the green region of the spectrum ( $\lambda = 555 \text{ nm}$ ). This is confirmed by the curve of the relative spectral sensitivity for the eye, plotted in Sec. 8.1 for the day (cone-type) vision. On the other hand, the sensitivity maximum for rods is shifted towards the region of shorter waves and corresponds to about  $\lambda = 510 \text{ nm}$ . Accordingly, for a strong illuminance when the "day-time mechanism" is in operation, red shades will seem brighter than blue ones. For a weak illuminance with the light of the same spectral composition, blue shades may seem brighter since the "twilight mechanism" (of rods) operates in these conditions. For example, a red poppy seems to be brighter than a blue cornflower in daylight and, on the contrary, looks darker with weak illumination in the evening.

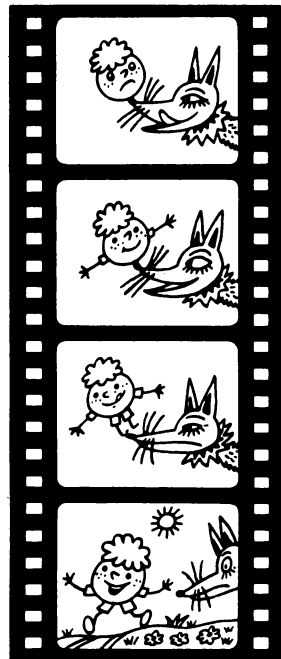
### 21.12. Duration of Visual Sensation

A decomposed substance stimulates the optic nerve during a short time of about  $1/7$  second. Therefore, an emerging visual sensation is retained during this time, although the stimulation itself may last for a shorter period. This ability of the eye to retain the gained impression during the above-indicated time is used in various appliances. The most known from them is cinematograph. On a cinema screen, a series of frames (Fig. 338) corresponding to consecutive positions of an object rapidly change one another (at a frequency on 24 frames per second). The eye retains the previous image at the moment when it starts to receive the next pattern. As a result, the perception of continuously changing positions of the object produces the impression of a smooth motion.

In order to obtain a film, a moving object must be photographed consecutively at the same frequency as that of projecting of the sequence of photographs on the screen, i.e. with 24 frames per second. If the rate of projecting is higher or lower than the rate of shooting, the observed picture will be distorted on the time scale. This effect is used for scientific purposes. By shooting at a very high frequency (say, 2000 frames per second) and projecting the film at a frequency of 20 frames per second, we extend a phenomenon in time (by a factor of 100), i.e. observe it at a slackened pace. This allows one to examine the details of fast processes ("time magnifier"). On the contrary, by shooting a slow process (like the growth of a crystal) at a very low frequency and projecting the film at a very high rate, we can reproduce the process in an accelerated pace and make visual the processes which change too slowly. This is used, for example, for reproducing the eruption of solar protuberances (by shooting them at a rate 500-600 times higher than the rate of projecting).

Fig. 338.

A fragment of a film. During a rapid movement of the frames, the impression of continuously changing positions (movement) is produced.





?

**III.1.** Derive the law of reflection of light with the help of Huygens' principle.

**III.2.** Figure 339 shows the arrangement of maxima on an interference pattern for  $\lambda = 400$  nm. Prove that for  $\lambda = 800$  nm, lines  $ab$ ,  $nn$ ,  $n'n'$ ,  $qq$  and  $q'q'$  will, as before, correspond to the positions of maxima, while lines  $mm$ ,  $m'm'$ ,  $pp$  and  $p'p'$  will give the positions of minima.

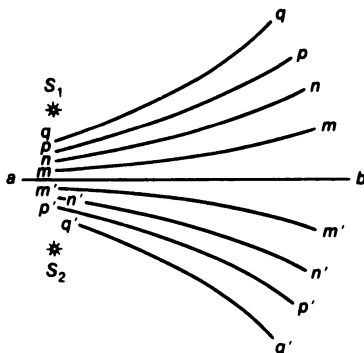


Fig. 339.

To Exercise III.2:  $S_1$  and  $S_2$  — the positions of coherent light sources,  $ab$  — the symmetry axis,  $mm$ ,  $m'm'$ ,  $nn$ ,  $n'n'$ ,  $pp$ ,  $p'p'$ ,  $qq$  and  $q'q'$  — the positions of maxima for  $\lambda = 400$  nm.

**III.3.** The path difference for rays in thin films is equal to  $2h$  for transmitted light and to  $2h + \lambda/2$  for reflected light, where  $h$  is the film thickness and  $\lambda$  is the wavelength. Prove that the radii of bright Newton's rings are proportional to the square roots of even numbers for transmitted light, and the radii of the dark rings are proportional to the square roots of odd numbers for this light, while all is the other way round for reflected light.<sup>7</sup>

**III.4.** For experiments with Newton's rings, a plano-convex lens with a radius of curvature of 10 m is used. (a) Determine the radius of the tenth dark ring for transmitted and reflected yellow light ( $\lambda = 600$  nm). (b) Determine the wavelength of the green line of mercury if it produces the second bright ring of radius 2.862 mm for reflected light. (c) Determine the separation between the second dark Newton's rings for reflected light, which correspond to two yellow lines of sodium:  $\lambda_1 = 589.0$  nm and  $\lambda_2 = 589.6$  nm. (d) Which dark ring for the reflected green light of the line of copper ( $\lambda = 515$  nm) has a radius of 6 mm?

**III.5.** What is the radius of curvature of the lens in Newton's experiment if the red line of hydrogen ( $\lambda = 656$  nm) produces the eighth bright ring of radius 8.6 mm?

**III.6.** Observing Newton's rings for the yellow light of a sodium line, Fizeau discovered that the sharpness of the pattern gradually decreased with increasing number  $N$  of the ring. For  $N = 500$ , the interference pattern was completely blurred, i.e. no sharp maxima alternating with minima were observed. Upon a transition to larger rings ( $N > 500$ ), the sharpness improved again.

This phenomenon can be explained as follows. The yellow light of sodium corresponds to two close lines  $\lambda_1$  and  $\lambda_2$ . It is known that  $\lambda_1 = 589.0$  nm. Determine  $\lambda_2$  from the observations described above. Which number  $N > 500$  corresponds to the next sharpness?

<sup>7</sup> The relations derived while solving this problem can be used in Exercises III.4 and III.5.

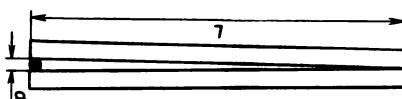


Fig. 340.

To Exercise III.7.

**III.7.** A wire having a diameter  $d = 10 \mu\text{m}$  is pressed between two glass plates (Fig. 340) so that an air wedge is formed. The length  $L$  of a plate is 10 cm. What will be the form of the interference pattern? What will be the separation between two neighbouring dark fringes if the plates are illuminated by the green light from a mercury lamp ( $\lambda = 540 \text{ nm}$ )? How will the width of the fringes (viz. the separation between two adjacent maxima) change when the angle between the plates increases (as a result of an increase in  $d$  or a decrease in  $L$ )?

**III.8.** Using the results of Ex. III.7, explain why the interference fringes in the pattern shown in Fig. 266 become narrower in the lower part of the film.

**III.9.** It is known that  $d = 20 \mu\text{m}$  and  $\lambda = 500 \text{ nm}$  in the experiment represented in Fig. 340. How many interference fringes will fit into the surface of the glass plate? What will be the dependence of the number of fringes on the gap width  $d$ ? How will the number of fringes depend on the size of the plate?

**III.10.** Two coherent sources  $S_1$  and  $S_2$  are separated by a distance  $l$  from each other. Interference fringes are observed on a screen placed at a distance  $D$  from the sources (Fig. 341). Calculate the width of an interference fringe, i.e. the distance  $h$  between two adjacent maxima, if the wavelength is  $\lambda$ . The distance  $D$  is large in comparison with  $l$  and  $\lambda$ . The positions of maxima on the screen correspond to the points for which the difference in the distances to  $S_1$  and  $S_2$  is equal to an integral number of wavelengths.

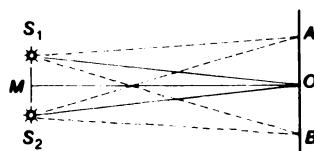


Fig. 341.

To Exercise III.10:  $S_1S_2 = l$ ,  $MO = D$  and  $OA = h$ .

**III.11.** A point light source  $S$  is placed in front of double prism (biprism) whose obtuse angle is close to  $180^\circ$ . Prove that the beams refracted by both halves of the prism interfere as if they were emitted by two coherent sources  $S_1$  and  $S_2$  (Fig. 342).

Calculate the distance  $S_1S_2$  between these coherent sources if the obtuse angle of the biprism is  $179.8^\circ$ . The distance  $SB$  from  $S$  to the biprism is 10 cm and the refractive index of the glass of which the biprism is made is 1.5. Pay attention to the fact that angles  $CAB$  and  $ACB$  of the prism are very small.

**III.12.** A narrow slit parallel to the edge of the biprism and illuminated by the yellow light of sodium ( $\lambda = 589 \text{ nm}$ ) is used as source  $S$  from the previous problem. The interference pattern is observed on a screen located at 10 m from  $S$ . Prove that the central maximum of the interference pattern lies at the point where the extension of line  $SB$  intersects

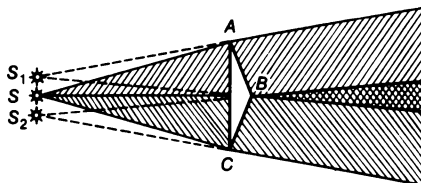


Fig. 342.

To Exercise III.11: for the sake of clarity, angles  $A$  and  $C$  of the biprism are exaggerated, the beams of rays incident on the lower and upper halves of the biprism are hatched differently.

the screen (see Fig. 342). Determine the positions of other maxima and minima on the screen. Calculate the width of an interference fringe, i.e. the distance between adjacent maxima (or minima). How will it vary with (a) decreasing obtuse angle of the biprism? (b) increasing distance to the screen?

**III.13.** It is shown in Exs. III.10 and III.11 that the width of interference fringes is the larger, the smaller the separation between the two coherent sources.

The interference in the light reflected from a thin film can be regarded as the interference of light from two coherent sources which are the reflections of a light source at the upper and lower surface of the film. How will the width of the fringes change if the film becomes thicker?

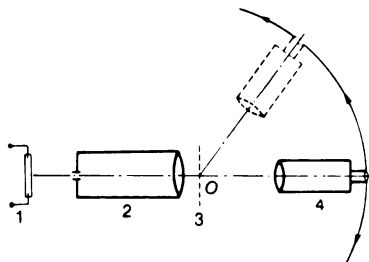
**III.14.** Derive the formulas for the first and second Fresnel zones for a point separated by a distance  $D$  from the front of a plane wave of wavelength  $\lambda$ .

**III.15.** Calculate the areas of the first, second and third Fresnel zones for a point separated by 2 m from the front of a plane wave whose wavelength is 500 nm.

**III.16.** What is the wavelength corresponding to the third-order maximum of a diffraction grating and coinciding with the fourth-order maximum for the wavelength  $\lambda = 405$  nm?

**III.17.** For what wavelengths can diffraction maxima be observed with a grating whose period is  $d$ ?

**III.18.** A monochromatic light of wavelength  $\lambda$  is incident on a diffraction grating with the grating constant  $d$ . The spectra are examined through a telescope located as shown in Fig. 343. How many spectral orders can be observed? Obtain the general solution and apply it to a particular case when  $d = 0.01$  mm and  $\lambda = 520$  nm.



**Fig. 343.**

To Exercise III.18: 1 — source of monochromatic light, 2 — collimator, 3 — diffraction grating, 4 — tube that can be rotated about centre  $O$ .

**III.19.** How many lines per millimetre must be in a diffraction grating suitable for investigating infrared spectra of a wavelength of about  $100\ \mu\text{m}$ ?

**III.20.** Derive the ratio of the wavelengths whose  $m$ th- and  $n$ th-order maxima coincide.

Solve the problem for a diffraction grating. (a) What are the wavelengths of the lines of the second- and third-order spectra that are superimposed on the line corresponding to the wavelength  $\lambda = 600$  nm of the first-order spectrum? (b) What is the wavelength of the lines of the first-order spectrum that are superimposed on the line of the wavelength  $\lambda = 450$  nm of the second-order spectrum?

**III.21.** A diffraction grating contains 100 lines per millimetre. Determine the angles at which the first-, second- and third-order maxima for  $\lambda = 500$  nm are located.

**III.22.** A grating spectroscope is designed as shown in Fig. 344. The grating constant is  $6\ \mu\text{m}$  and the focal length of objective 3 is 1 m. (a) Determine the separation between two yellow lines of 589.0 nm and 589.6 nm for sodium in the first- and second-order spectra. (b) Determine the distance between the positions of the line of 600 nm in the first- and second-order spectra. (c) In which order is the separation between two yellow lines of 577 nm and 579 nm for mercury equal to 1.33 mm? (d) The dispersion of the spectroscope is measured by the number of nanometres corresponding to the region of the plate 1 mm long. Does the dispersion of the grating spectroscope depend on the wavelength? Calculate the dispersion of the spectroscope for the first- and second-order spectra.

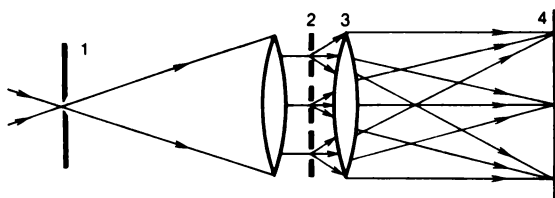


Fig. 344.

To Exercise III.22: 1 — collimator, 2 — diffraction grating, 3 — objective of the camera, 4 — photographic plate.

**III.23.** If the filament of an incandescent lamp is viewed by an observer screwing up his eye, it seems to be bordered by light patches in two perpendicular directions. If the observer turns his head about the direction of view, the pattern rotates as well. If the filament is almost parallel to the observer's nose, a sequence of coloured (iridescent) images of the filament can be seen. However, this is impossible or nearly possible when the filament is at right angles to the observer's nose.

Make this experiment.

Pay attention to the sequence of colours in the coloured image.

Explain the observed phenomena.

**III.24.** Assuming that the thickness of an eyelash in the previous problem is 0.1 mm and considering that the eyelashes are separated by 0.15 mm, calculate the approximate distance between the images of the filament if a lamp is at 3 m from the observer. Will this distance change as a result of moving the lamp closer to or away from the observer? Verify this result experimentally.

**III.25.** A more accurate schematic diagram of experiment on determining the velocity of light by Foucault's method is shown in Fig. 345.

A lens produces an image of source  $S$  on the surface of a spherical mirror whose centre coincides with the rotational axis. A glass plate reflecting a fraction of the light beam in direction  $S''$  facilitates observations. Analyze the operation of the set-up.

**III.26.** The resolving power of a telescope is such that two stars with an angular separation of  $1/8''$  are distinguished by the instrument as separate objects. What must be the distance (in km) between such distinguishable stars if the light from them reaches the Earth in 100 light years?

**III.27.** The resolving power of the eye is  $1'$  for a sufficiently high illuminance. Thin black wires are stretched against a white background at a distance of 1 m from the eye. What must be the separation between the wires for the eye to be able to distinguish them?

**III.28.** Why is a short-sighted eye able to distinguish smaller details (say, read smaller print) than a normal eye?

**III.29.** The diameter of a microscope objective is close to that of the eye's pupil. Therefore, their angular resolving powers due to the diffraction at the aperture of the pupil or of the objective are approximately the same and equal to  $1'$ . But since the focal length

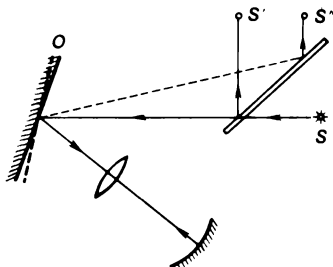


Fig. 345.

To Exercise III.25.

of the objective is small, the object under investigation can be brought closer to the lens. What must be the separation between the lines of a grating so that they can be distinguished through a microscope with a focal length of the objective equal to 1 mm?

**III.30.** The resolving power of the eye (sharpness of vision) depends on the illuminance and the nature of an observed object. A normal eye may distinguish the details of a white object against a dark background (say, the letters written in a chalk on a blackboard) if the angular size of the letters is about  $100''$  (about  $2'$ ) at an illuminance of about 100 lx. What must be the size of the letters on the blackboard for a pupil to be able to distinguish them while sitting at his desk located at 8 m from the blackboard? The details by which a letter is distinguished from others are of about one-fifth of the letter size.

**III.31.** What must be the distance between two points on the Moon (say, two peaks of the mountains) so that they can be distinguished by the unaided eye and through a telescope? The illuminance and the contrast are assumed to be sufficiently high to consider that the resolving power is  $1'$  for the eye and  $1/8''$  for the telescope. The distance from the Earth to the Moon is 382 000 km.

**III.32.** A sheet of white paper is illuminated simultaneously by two electric arcs screened by a yellow and a blue glass respectively (Fig. 346a). The yellow glass absorbs the blue, indigo and violet regions of the spectrum, and the blue glass, the red, orange and yellow regions.

The same sheet of paper brightly illuminated by the electric arc is viewed through the same two glasses piled together (Fig. 346b). Explain the colours of the illuminated paper in the former and latter cases.

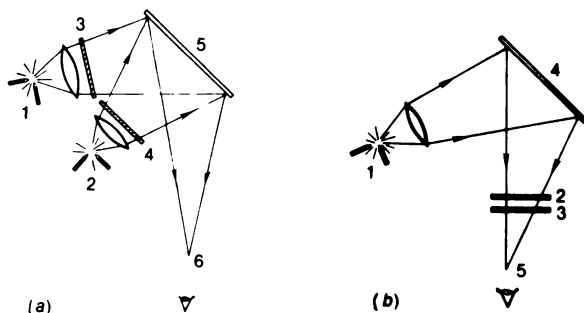


Fig. 346.

To Exercise III.32: (a) 1 and 2 — arcs, 3 and 4 — yellow and blue glasses, 5 — white paper, 6 — eye, (b) 1 — arc, 2 and 3 — yellow and blue glasses, 4 — white paper, 5 — eye.

**III.33.** Describe the colours of white, red, yellow, green and blue papers illuminated by the yellow light of sodium flame.

**III.34.** Explain the origin of the colour of (a) blue sky; (b) blue glass and (c) blue paper.

**III.35.** An ultraviolet radiation of a wavelength of 200 nm is incident on a nickel plate characterized by a work function of 4.5 eV. Determine the maximum velocity of electrons emitted as a result of photoelectric effect.

The values of the required constants are as follows: the electron mass is  $0.91 \times 10^{-30}$  kg, the speed of light is  $3 \times 10^8$  m/s and Planck's constant is  $6.6 \times 10^{-34}$  J·s.

**III.36.** What is the maximum wavelength of the light causing the photoelectric effect from the surface of sodium (the work function  $A_{Na} = 2.35$  eV), tungsten ( $A_W = 4.5$  eV) and platinum ( $A_{Pt} = 5.3$  eV)? (This wavelength is known as the *photoelectric threshold*.)

**III.37.** A zinc plate (see Fig. 330) exposed to X-rays has been charged so that the electrometer indicates 1500 V. (1) What will be the sign of the charge on the electrometer? (2) What will be the wavelength of the X-rays used in this experiment? (3) Will the result of this experiment change noticeably if the plate is made of nickel or tungsten?

**III.38.** Calculate the ratio of the path lengths of solar rays in the atmosphere for the positions of the Sun at the horizon and at the zenith (cf. Fig. 319).

Assume that the density of the atmosphere is uniform and the same as at the surface of the Earth (the so-called reduced atmosphere). Its thickness should be taken as 10 km and the radius of the Earth as 6400 km.

**III.39.** It sometimes happens that a blackboard "gleams", i.e. the letters written on it in white chalk cannot be distinguished on the blackboard. Explain this phenomenon. At what relative positions of pupils, the blackboard and a window will this be observed? Will a screen made of black velvet gleam?

*Remark.* The letters written in chalk reflect (scatter) light diffusely and have a large reflection coefficient (the albedo for chalk is close to unity). The polished blackboard exhibits the mirror reflection, although it is characterized by a small reflection coefficient. This coefficient increases noticeably when the angle of incidence of light on the blackboard approaches  $90^\circ$ .

**III.40.** Given two optical filters: a violet and a yellow-green. The former transmits the violet and dark-blue parts of the spectrum, while the latter transmits red, orange, yellow and yellow-green parts. Piled together, they consequently absorb all the colours of the spectrum. Such filters are known as *compensating*.

The light from an electric arc is directed to a white sheet of paper or a vessel containing fluorescein. The filters are located in one of the four positions shown in Fig. 347.

What will be observed in the former (paper) and latter (fluorescein) case?

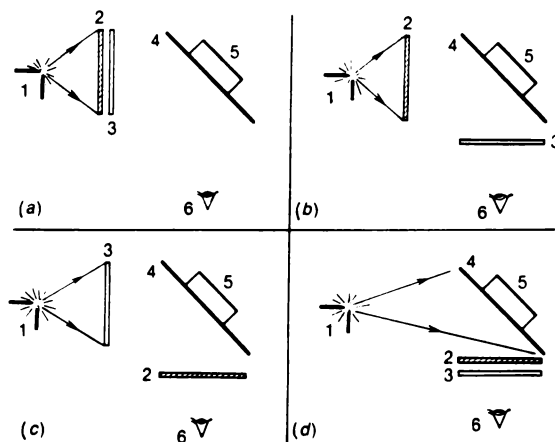


Fig. 347.

To Exercise III.40: 1 — arc, 2 — violet filter, 3 — yellow-green filter, 4 — sheet of paper, 5 — vessel with fluorescein, 6 — eye.

## Part Four

# Atomic and Nuclear Physics

## Chapter 22

# Atomic Structure

### 22.1. Atoms

It is well known from chemistry and previous parts of physics that all bodies consist of individual very small particles, viz. atoms and molecules. An *atom*<sup>1</sup> is the smallest particle of a given chemical element. A *molecule* is a compound particle consisting of several atoms. The physical and chemical properties of elements are determined by the properties of atoms of their elements.

Up to the end of the 19th century it was believed that atoms are the simplest indivisible particles of matter. However, the subsequent development of science refuted this opinion. It was established that atoms are not elementary particles but are rather quite complicated objects. This follows, for example, from optics, in particular, from the electromagnetic theory of light. It has been proved that electromagnetic waves, and hence light as well, are emitted during an accelerated motion of electric charges. But atoms of a substance can also emit light, viz. visible electromagnetic waves, producing an emission spectrum typical of each atom (see Chap. 20). Hence we may conclude that atoms contain electric charges that can move. The analysis of electrical conductivity of metals and gases (see Vol. 2, Secs. 7.1, 8.1 and 8.2) indicates that atoms contain negatively charged particles, viz. electrons, whose mass is very small in comparison with the atomic mass. Since an atom is neutral as a whole, it must contain, in addition to electrons, positively charged particles as well.

Thus, atoms are composite particles built from other, simpler particles. The constituents of atoms are *electrons* and, as will be shown in Chap. 24, positive particles (*protons*) and neutral particles (*neutrons*).

Atoms are rather strong systems which are much more stable than molecules composed of several atoms. Indeed, a molecule can be decomposed into atoms without any difficulty. For this purpose, it is sufficient, for example, to heat a substance. For instance, by heating gaseous nitrogen or hydrogen to about 2000 K, we make a considerable fraction of molecules ( $N_2$  or  $H_2$ ) decompose into corresponding atoms. It should be noted here

---

<sup>1</sup> The word “atom” takes its origin from the Greek *atomos* meaning indivisible.

that the molecules  $N_2$  and  $H_2$  are among the strongest molecules. A molecule of ammonium chloride  $NH_4Cl$ , for example, is decomposed into ammonia  $NH_3$  and hydrogen chloride  $HCl$  even at room temperature or as a result of slight heating. Dropping a piece of sodium metal into water, we cause a chemical reaction as a result of which water molecules disintegrate, gaseous hydrogen  $H_2$  is liberated and sodium hydroxide  $NaOH$  is formed. In other words, a radical molecular transformation takes place. Similar transformations of atoms could not be realized for a long time. Quite strong effects (like heating, a change in pressure, or the passage of powerful electric discharges) lead only to insignificant changes in atoms: they can be ionized, i.e. one or several electrons can be detached from them.

Although an ion has some features which distinguish it from its atom, it preserves the main properties of the atom. The ion easily becomes a neutral atom again by capturing electrons missing in the process of ionization. Prolonged efforts of alchemists to convert one atom into another (in particular, to obtain gold from "unnoble" metals) via various chemical or physical transformations remained fruitless.

Phenomena in which atoms undergo deep changes and are converted into the atoms of other elements have been discovered comparatively recently. These phenomena will be described in Chaps. 23 and 24.

## 22.2. Avogadro's Constant. Size and Mass of Atoms

One of the most important constants in atomic physics is *Avogadro's constant* (see Vol. 1, Sec. 13.21), which indicates the number of structural elements (atoms, molecules or ions) in a mole of a substance. Knowing Avogadro's constant, we can calculate the quantities that characterize an individual atom: its mass and size, charge of an ion, etc.

There are several methods for measuring Avogadro's constant. They are based on various types of physical phenomena such as Brownian movement of particles suspended in a fluid (see Vol. 1, Sec. 12.7), radioactivity (see Chap. 22) and scattering of light in gases. The method based on the diffraction of X-rays is among the most accurate methods of determining this constant.

It is known from optics (see Chap. 17) that X-rays are electromagnetic waves differing from visible light in that they have much smaller wavelength. The wave nature of X-rays was established for the first time in experiments on diffraction in crystals. These experiments simultaneously confirmed the correctness of the idea about crystals as the systems of regularly arranged atoms forming a space lattice (see Vol. 1, Sec. 15.4).

A beam of X-rays incident on a crystal is scattered mainly along a few selected directions (see Sec. 17.5). The scattering angles are determined by



the wavelength of X-rays and by the separation between adjacent atoms in the crystal. If one of these quantities is known, the other can be determined from the measured values of scattering angles.

The wavelength of X-rays is measured with a high accuracy on the basis of diffraction at an ordinary ruled diffraction grating used in optics (see Secs. 14.8 and 14.11). But if we know the wavelength of X-rays, we can determine the interatomic distance in a crystal. In the crystals of NaCl type, the atoms are arranged at the vertices of a cube with an edge equal to the shortest interatomic distance  $a$ .<sup>2</sup> The volume of the crystal per atom is  $a^3$ , and per molecule,  $2a^3$ . Let the volume of a mole of a crystalline substance be  $V$ . Then Avogadro's constant can be calculated from the formula

$$N_A = \frac{V}{2a^3}.$$

All known methods of measuring Avogadro's constant lead to the same value. According to latest measurement, this value is

$$N_A = 6.02 \times 10^{23} \text{ mole}^{-1}.$$

The coincidence of the results of measurements of Avogadro's constant (as well as the coincidence of the values of masses, sizes and velocities of atoms measured by different methods) is a convincing proof of the validity of the theory of atomic structure.

Let us consider the sharp difference in the compressibility of gases on the one hand and of liquids and solids on the other hand.

According to Boyle's law (see Vol. 1, Sec. 8.6), in order to reduce the volume of a gas by 1%, it is sufficient to increase its pressure by 1%. In liquids and solids, however, the same increase in volume (by 1%) requires an increase in pressure by tens and hundreds of times (it is assumed that the initial pressure is equal to the atmospheric pressure). This difference is due to the fact that gas molecules are located at distances much longer than the size of a molecule. The thermal motion prevents them from coming closer together. The forces of interaction between gas molecules separated by such large distances are so weak that they can be neglected. On the contrary, for liquids and solids we can assume that their atoms (or molecules) are packed closely. When atoms (or molecules) are brought still closer, huge repulsive forces emerge, which hamper the reduction of the volume of such bodies.

Thus, the mean distance between the centres of adjacent atoms of a solid or a liquid can be approximately regarded as the linear atomic size. Knowing Avogadro's constant, we can easily calculate this distance.

---

<sup>2</sup> The arrangement of individual atoms in heavy elements in a crystal can be observed and their separation can also be measured with the help of an electron microscope.

A mole of a substance contains  $N_A$  atoms and occupies a volume of  $M/\rho$ , where  $\rho$  is the density of the substance and  $M$  is its molar mass. Let us take a mole of the substance in the form of a cube. The number of atoms fitting into the edge of this cube is  $\sqrt[3]{N_A}$ . The length of this edge is equal to the cube root of its volume, i.e.  $\sqrt[3]{M/\rho}$ . Dividing the length of the edge by the number of atoms fitting into it, we obtain the mean distance between the centres of adjacent atoms, which is regarded as the approximate size of an atom. This distance is

$$a = \frac{\sqrt[3]{M/\rho}}{\sqrt[3]{N_A}} = \sqrt[3]{\frac{M}{\rho N_A}} \quad \checkmark$$

For liquid hydrogen (at 24 K), we substitute  $M = 1.008 \times 10^{-3}$  kg/mole and  $\rho = 86 \text{ kg/m}^3$  and obtain

$$a = \sqrt[3]{\frac{1.008 \times 10^{-3}}{86 \times 6.02 \times 10^{23}}} = 2.7 \times 10^{-10} \text{ m.}$$

For other elements, the calculation gives similar results. We can conclude that the linear sizes of all atoms are close to  $10^{-10}$  m.

Knowing Avogadro's constant, we can also determine the mass of an atom:  $m = M/N_A$ . ✓

This formula gives us the mean value of the atomic mass. The question of whether or not all atoms of a given element have the same mass must be answered experimentally (see Sec. 22.5).

The lightest of all atoms is a hydrogen atom whose relative atomic mass is 1.008, and hence  $M = 1.008 \times 10^{-3}$  kg/mole. Dividing this value of  $M$  by  $N_A$ , we obtain the mass of the hydrogen atom:

$$m_H = \frac{1.008 \times 10^{-3}}{6.02 \times 10^{23}} = 1.67 \times 10^{-27} \text{ kg.}$$

### 22.3. Elementary Electric Charge

The laws of electrolysis discovered by Faraday justify the existence of the smallest indivisible amounts of electricity. In electrolysis, a mole of any  $n$ -valent element carries a charge  $F \cdot n = 96487n$  coulombs ( $F$  is Faraday's constant). Thus, the charge corresponding to an atom (to be more precise, to an ion)

$$q = \frac{Fn}{N_A} = \frac{96480}{6.02 \times 10^{23}} n = 1.60 \times 10^{-19} n \text{ C.}$$

A monovalent ion ( $n = 1$ ) corresponds to the charge  $e = 1.60 \times 10^{-19}$  C, a bivalent ion ( $n = 2$ ) is carried by a charge of  $2e$ , a trivalent ion ( $n = 3$ ) has a charge of  $3e$ , and so on.

This regularity can easily be explained if we assume that the charge  $e = 1.60 \times 10^{-19} \text{ C}$  is the minimum portion of charge, or *elementary charge*.

However, the laws of electrolysis can be interpreted so that  $e$  is the average charge carried by a monovalent ion. The property of an  $n$ -valent ion to carry an  $n$  times larger charge should then be explained by the properties of the ion rather than by the atomistic nature of electricity. For this reason, direct experiments were required in which the smallest amount of electricity could be measured in order to verify the existence of an elementary charge. Such experiments were carried out in 1909 by the American physicist Robert Andrews Millikan (1868-1953).

The set-up for Millikan's experiment is shown schematically in Fig. 348. Its main part is a parallel-plate capacitor 2, 3. A potential difference of either sign can be applied to the plates through switch 4. An atomizer sprays tiny drops of oil or other liquid into vessel 1. Some of these drops get through a hole in the upper plate into the space between the capacitor plates, which is illuminated by lamp 6. The drops are observed in a microscope through window 5. They look like bright stars against a dark background.

In the absence of an electric field between the capacitor plates, drops fall down at a constant velocity. When the field is switched on, neutral drops continue to fall at the constant velocity. Many drops, however, acquire a charge during spraying (electrostatic charging by friction). In addition to the force of gravity, such charged drops experience the action of the force exerted by the electric field. Depending on the sign of the charge, the direction of the field can be chosen so that the electric force is directed

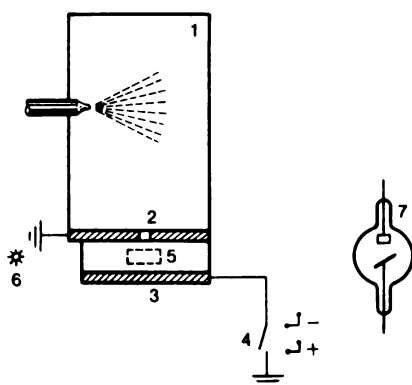


Fig. 348.

Schematic diagram of the experiment on measuring the elementary electric charge. X-ray tube 7 is intended for measuring the charge of the drops; the radiation emitted by it produces in the volume between plates 2 and 3 ions which stick to a drop and change its charge.

against the force of gravity. Then a charged drop will fall in the electric field at a lower velocity than in the absence of the field. The magnitude  $E$  of the electric field can be chosen so that the electric force becomes stronger than the force of gravity, and the drop starts to move upwards.

Millikan's device makes it possible to watch the same drop for a few hours. For this purpose, it is sufficient to switch off (or reduce) the field at the moment the drop approaches the upper plate of the capacitor and switch it on (or increase) again when the drop reaches the lower plate.

If a drop moves uniformly in the absence of a field, this means that the force exerted on it is balanced by the air resistance which is proportional to the velocity of the drop. Therefore, we can write the following equation for such a drop:

$$mg = kv, \quad (22.3.1)$$

where  $mg$  is the force of gravity acting on the drop of mass  $m$ ,  $v$  is the velocity of the drop,  $kv$  is the air drag (force of friction) and  $k$  is the coefficient determined by the viscosity of air and the size of the drop.

Having measured the diameter of the drop with the help of the microscope (and hence knowing its mass) and having then determined the velocity  $v$  of the uniform free motion, we obtain from (22.3.1) the value of the coefficient  $k$  which remains unchanged for the given drop.

The condition of a uniform ascent at a velocity  $v_E$  of a drop with a charge of  $q$  in an electric field  $E$  has the form

$$qE - mg = kv_E. \quad (22.3.2)$$

Hence

$$q = \frac{mg + kv_E}{E}.$$

Thus, making measurements for the same drop in the absence of the field and with it, we can determine the charge  $q$  of the drop. This charge can be varied. For this purpose, an X-ray tube 7 can be used (see Fig. 348) to ionize the air in the capacitor. Formed ions will be captured by the drop, and its charge will be changed ( $q'$ ). This will bring about a change in the velocity of the uniform motion of the drop (it becomes  $v'_E$ ) so that

$$q'E - mg = kv'_E. \quad (22.3.3)$$

From (22.3.2) and (22.3.3), we obtain the change in charge:

$$q - q' = \frac{k}{E} (v_E - v'_E),$$

which can also be determined since the coefficient  $k$  for the drop is known.

Numerous measurements of this type with drops of different substances (water, oil, glycerine and mercury) bearing positive and negative charges revealed that charge  $q$ , as well as the observed change in charge  $q - q'$ , are always multiple to the same minimum charge

$$e = 1.60 \times 10^{-19} \text{ C.}$$

This minimum charge is equal to the elementary charge established for the process of electrolysis. It should be emphasized here that the initial charge of a drop is due to electrostatic charging by friction, while the change in it may occur due to capture by the drop of gas ions produced by X-rays. Thus, the charge formed by electrostatic charging, the charges of gas ions and the electrolyte ions are composed of the same elementary charges. The results of other experiments make it possible to generalize this conclusion: *all positive and negative charges encountered in nature consist of integral numbers of elementary charges  $e = 1.60 \times 10^{-19} \text{ C.}$*

In particular, the charge of an electron is equal in magnitude to an elementary charge.

#### 22.4. Units of Charge, Mass and Energy in Atomic Physics

Thus, the charge of any particle always contains an integral number of elementary charges. For an atomic particle, this number is not large. For this reason, in atomic physics it is convenient to use the unit of electric charge equal to the elementary charge  $e = 1.60 \times 10^{-19} \text{ C.}$  In atomic physics the unit of mass is equal to 1/12 of the atomic mass of the carbon isotope  $^{12}\text{C}$ . The atomic mass  $M$  of this isotope is equal to  $12 \times 10^{-3} \text{ kg/mole}$ . Therefore, the atomic mass unit (amu) is

$$1 \text{ amu} = \frac{1}{12} \frac{M}{N_A} = \frac{1}{12} \frac{12 \times 10^{-3}}{6.02 \times 10^{23}} \text{ kg} = 1.66 \times 10^{-27} \text{ kg.}$$

The atomic mass unit can also be defined as the mass of an element whose atomic mass is unity. Therefore, the mass of the atom (to be more precise, its mean value) in atomic mass units is equal to the atomic mass of this element.

It should be noted that an element with an atomic mass equal to unity does not exist in nature. The atomic mass of hydrogen is close to unity but is somewhat larger (it is equal to 1.008). Therefore, the mass of the lightest atom is 1.008 amu.

[The unit of energy adopted in atomic physics is the energy acquired by a particle with a charge of  $e$  (say, an electron) as a result of its passage across a potential difference of 1 V.] This unit is termed an *electron volt* and denoted by eV. The energy acquired by a charged particle moving in

an electric field is equal to product of the charge and the potential difference between the initial and final points of the path. Therefore,

$$1 \text{ eV} \approx 1.6 \times 10^{-19} \text{ C} \cdot 1 \text{ V} \approx 1.6 \times 10^{-19} \text{ J}.$$

It follows from the definition of an electron volt that an electron accelerated by the potential difference  $U$  [V] has an energy numerically equal to  $U$  [eV]. Having passed through the same potential difference, an ion with a charge of  $2e$  acquires an energy of  $2U$  [eV], and so on.

The energy of not only charged but also neutral particles can be measured in electron volts. By way of an example, let us express in electron volts the energy of an oxygen atom ( $m = 16 \text{ amu}$ ) moving at a velocity  $v = 10^3 \text{ m/s}$ :

$$\begin{aligned} W &= \frac{mv^2}{2} = \frac{16 \times 1.66 \times 10^{-27} \times (10^3)^2}{2} \\ &= 1.33 \times 10^{-20} \text{ J} = \frac{1.33 \times 10^{-20}}{1.6 \times 10^{-19}} \text{ eV} = 0.083 \text{ eV}. \end{aligned}$$

Units multiple to electron volt are also used:

$$\begin{array}{ll} 1 \text{ keV} = 10^3 \text{ eV}, & 1 \text{ MeV} = 10^6 \text{ eV}, \\ 1 \text{ GeV} = 10^9 \text{ eV}, & 1 \text{ TeV} = 10^{12} \text{ eV}. \end{array} \quad \checkmark$$

## 22.5. Measurement of Mass of Charged Particles.

### Mass Spectrograph

From the course on electricity, it is known that a charged particle moving in a magnetic field is acted upon by a force known as the *Lorentz force*.<sup>3</sup> The Lorentz force is perpendicular to the magnetic field and to the velocity of the particle, its direction being determined by the left-hand rule (Fig. 349). The magnitude of this force is proportional to the charge  $q$  of the particle, its velocity  $v$ , the magnetic induction  $B$  of the field and the sine of the angle between vectors  $v$  and  $B$ . If the velocity  $v$  is at right angles to the magnetic induction  $B$ , the magnitude of the Lorentz force is given by

$$F_L = qvB,$$

where  $q$  is the charge of the particle in coulombs,  $v$  is its velocity in metres per second,  $B$  is the magnetic induction in teslas and  $F_L$  is the Lorentz force in newtons. The acceleration  $a$  imparted by the Lorentz force, just like by any force in general, is directly proportional to the force and inversely proportional to the mass  $m$  of the particle.

<sup>3</sup> Most modern textbooks give a somewhat different definition of the Lorentz force, which includes the term  $qE$  (see Vol. 2, Sec. 15.1).

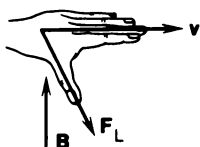


Fig. 349.

Direction of the Lorentz force  $F_L$  acting on a charge moving in a magnetic field of induction  $B$  at velocity  $v$ . In this case, we have a positive charge. For a negative charge, the force will have the opposite direction.

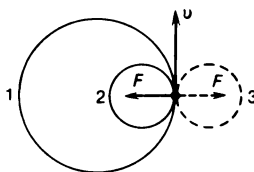
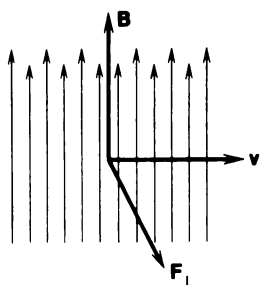


Fig. 350.

Trajectories of charged particles having equal initial velocities in a uniform magnetic field: 1 — small  $q/m$  ratio, 2 — large  $q/m$  ratio, 1 and 2 — negative particles, 3 — positive particle. The magnetic field lines are perpendicular to the plane of the figure and directed upwards.

Let us consider the motion of a particle in a uniform magnetic field directed at right angles to the velocity of the particle. Since the Lorentz force, and hence the acceleration, is perpendicular to the velocity, the particle will move in a circle so that the magnitude  $v$  of its velocity remains unchanged since, as is well known from mechanics, the velocity and acceleration are at right angles for the uniform motion in a circle. The acceleration of a particle moving uniformly in a circle is  $v^2/r$ , where  $r$  is the radius of the circle. Thus, we can write for the acceleration

$$a = \frac{q}{m} v B = \frac{v^2}{r},$$

whence

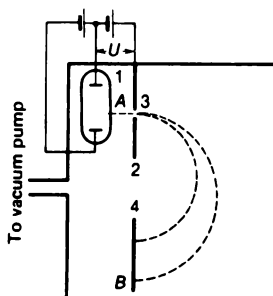
$$r = \frac{mv}{qB}. \quad (22.5.1)$$

The smaller the ratio  $q/m$ , the larger the radius of the trajectory of the particle for given  $v$  and  $B$  (Fig. 350). Knowing  $v$  and  $B$  and having measured the radius  $r$  of the trajectory, we can determine  $q/m$ , viz. the charge-to-mass ratio for the particle. The particle's charge is equal to one or several elementary charges. If it is known, we can determine the mass of the particle. This principle underlies the operation of an instrument known as the *mass spectrograph* and intended for measuring the masses of the smallest charged particles, viz. ions and electrons.

Figure 351 shows the schematic diagram of a mass spectrograph with a uniform magnetic field. It consists of a vessel (in which a high vacuum is created) placed in a magnetic field whose lines are perpendicular to the

Fig. 351.

Schematic diagram of a mass spectrograph: 1 — source of ions (gas-discharge tube), 2 — diaphragm with a slit, 3 and 4 — photographic plate,  $U$  — voltage accelerating ions.



plane of the figure. Charged particles are emitted by source 1 (the simplest source is an electric discharge in a gas, which is accompanied by intensive ionization). Electrons and negative ions will be “pumped out” from the discharge when a positive potential difference is applied between diaphragm 2 and the slit of the source, and positive ions when the potential difference is negative. By filling the source with different gases and vapours, one can obtain ions of various elements.

Particles flying through slit 3 get into the magnetic field at the velocities imparted to them by the accelerating potential difference. All the particles having the same  $q/m$  ratio acquire identical velocities and describe the circles of the same radius in the magnetic field. The beam of particles deflected through  $180^\circ$  is incident on a photographic plate. A dark line will be seen on the plate after it has been developed.<sup>4</sup> Distance  $AB$  is equal to twice the radius  $r$  of the circle in which the particle moves (see Fig. 351). The magnitude of  $r$  depends on the particle velocity. In order to find the velocity, we make use of the fact that the particle enters the magnetic field when its kinetic energy is  $W_k = mv^2/2$ . This energy is due to the work done by the electric field and equal to  $qU$ . Therefore,

$$W = \frac{1}{2} mv^2 = qU. \quad (22.5.2)$$

From (22.5.1) and (22.5.2) we have

$$m = \frac{qr^2 B^2}{2U}.$$

Substituting the known values of  $q$ ,  $B$  and  $U$  and the radius  $r$  obtained as a result of measurements into this formula, we can calculate the mass of the particles incident at points  $B$  of the plate.

If the beam of particles emitted by the source contains particles with different charge-to-mass ratios, several parallel lines will be obtained on

<sup>4</sup> Experiments show that fast charged particles produce the same effect on a light-sensitive emulsion of photographic plates as light rays.



the plate. The line closest to the slit is due to the particles moving in a circle of the minimum radius. These particles have the maximum value of the charge-to-mass ratio. If all particles in the beam have the same charge, the line closest to the slit corresponds to the lightest particles.

By analogy with optics, the image formed on a photographic plate is called the *spectrum*. An optical spectrograph gives the spectrum of wavelengths of light beam, i.e. the distribution of spectral lines over wavelengths. A mass spectrograph gives the spectrum of masses of the beam particles, i.e. the mass distribution of particles (to be more precise, the  $q/m$  distribution).

## 22.6. Electron Mass.

### Velocity Dependence of Electron Mass

When the mass of an electron is measured in experiments with a mass spectrograph, only one line is formed on the photographic plate. Since the charge of every electron is equal to an elementary charge,<sup>5</sup> we arrive at the conclusion that all electrons have the same mass.

The electron mass, however, turns out to be variable. It increases with the potential difference  $U$  accelerating electrons in a mass spectrograph (see Fig. 351). Since the kinetic energy  $W_k$  of an electron is directly proportional to the accelerating potential difference  $U$  ( $W_k = eU$ ), the electron mass increases with its kinetic energy. Experiments lead to the following mass-energy relation:

$$m = m_0 + \frac{W_k}{c^2}, \quad (22.6.1)$$

where  $m$  is the mass of the electron having a kinetic energy  $W_k$ ,  $m_0$  is a constant and  $c$  is the speed of light in vacuum ( $c = 3 \times 10^8$  m/s). This formula shows that the mass of an electron at rest (i.e. an electron having a kinetic energy  $W_k = 0$ ) is  $m_0$ . This value is known as the *rest mass* of the electron.

Measurements carried out with different sources of electrons (gas discharge, thermionic emission, photoelectric emission, and so on) result in the identical value of the rest mass of an electron. This mass turns out to be extremely small:

$$m_0 = 0.911 \times 10^{-30} \text{ kg} = \frac{1}{1823} \text{ amu.} \quad \checkmark$$

**Thus, an electron (at rest or moving at low velocity) has a mass which is equal to about 1/2000 of the mass of the lightest substance (hydrogen).**

<sup>5</sup> Numerous experiments prove that electrons having two or more elementary charges do not exist.

The quantity  $W_k/c^2$  in relation (22.6.1) is an additional electron mass due to its motion. As long as this addition is small, while calculating the kinetic energy we can approximately take  $m_0$  for  $m$  and put  $W_k = m_0 v^2/2$ . Then  $W_k/c^2 = m_0 v^2/2c^2$ . It follows hence that our assumption about the smallness of the additional mass in comparison with the rest mass  $m_0$  is equivalent to the condition that the velocity of an electron is much smaller than the speed of light ( $v/c \ll 1$ ). On the contrary, when the velocity of an electron approaches the speed of light, the additional mass becomes large.

In 1905, Albert Einstein (1879-1955) theoretically substantiated formula (22.6.1) in the theory of relativity. He proved that it is applicable not only to electrons but also to all particles or bodies without exception, and  $m_0$  must be regarded as the rest mass of a particle of body under consideration. Einstein's conclusions were verified later in numerous experiments and fully confirmed. The theoretical mass-velocity relation derived by Einstein has the form

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}}. \quad (22.6.2)$$

Thus, the mass of any body increases with its kinetic energy or velocity. However, as in the case of an electron, the additional mass due to motion is noticeable only when the velocity of motion approaches the speed of light. Comparing (22.6.1) and (22.6.2), we obtain the formula for the kinetic energy taking into account the velocity dependence on mass:

$$W_k = m_0 c^2 \left( \frac{1}{\sqrt{1 - v^2/c^2}} - 1 \right). \quad (22.6.3)$$

In relativistic mechanics (i.e. mechanics based on the theory of relativity), as well as in classical mechanics, the momentum of a body is defined as the product of its mass and velocity. However, now the mass itself depends on velocity (see (22.6.2)), and the relativistic expression for momentum has the form

$$\mathbf{p} = m\mathbf{v} = \frac{m_0 \mathbf{v}}{\sqrt{1 - v^2/c^2}}. \quad (22.6.4)$$

In Newtonian mechanics, the mass of a body is assumed to be constant and independent of its motion. This means that Newtonian mechanics (to be more precise, Newton's second law) is applicable only to motions of bodies at velocities which are very low in comparison with the speed of light. The speed of light has a huge value so that the condition  $v/c \ll 1$  is always satisfied for motion of terrestrial or celestial bodies, and the mass of a body is practically the same as its rest mass. Expressions (22.6.3) and

(22.6.4) for kinetic energy and momentum at  $v/c \ll 1$  are transformed to the corresponding formulas in classical mechanics (see Ex. IV.11). Therefore, while considering the motion of such bodies, Newtonian mechanics can and must be used.

The situation is different in the world of the smallest particles, viz. electrons and atoms. Here we often encounter fast motions when the velocity of a particle is no longer small in comparison with the speed of light. In such cases, Newtonian mechanics is inapplicable, and more precise (although more complex) Einsteinian mechanics has to be employed. In this new mechanics, the dependence of the mass of a particle on its velocity (energy) is one of the basic relations.

Another typical conclusion of Einsteinian mechanics is that concerning the impossibility of motion with a velocity exceeding the speed of light in vacuum. The speed of light is the limiting velocity of motion of bodies.

The existence of the limiting velocity of bodies can be treated as a consequence of the increase in mass with velocity: the higher the velocity, the heavier the body, and the more difficult a further increase in velocity (since acceleration decreases with increasing mass).

## 22.7. Einstein's Law

In the previous section, we established the relation between the kinetic energy of a body and its mass: if a kinetic energy  $W_k$  is imparted to a body, its mass increases by  $W_k/c^2$ . This relation is of universal nature: it refers to all bodies, viz. large and small, charged and neutral, and so on. At the same time, kinetic energy is just one of many forms of energy. Other forms of energy known to us are the electric energy, the energy of light quanta, the internal energy of bodies, etc.

It is well known that all forms of energy can be converted into one another. A question arises: is the relation between all forms of energy and the mass of a body the same as for kinetic energy?

For one particular case, the answer to this question is affirmative. Let us suppose that we heat a monatomic gas. The increase in the internal energy of monatomic gases upon heating boils down to an increase in the kinetic energy of its particles.<sup>6</sup> But as was shown above, the mass of the particles increases with their kinetic energy. Consequently, the mass of the gas as a whole must increase upon heating. Since the body (gas) as a whole remains stationary (at rest), it follows hence that it is the rest mass that in-

---

<sup>6</sup> Strictly speaking, this refers to an ideal monatomic gas. But for real gases too the change in the potential energy of interaction of particles upon heating is negligibly small in comparison with the change in the kinetic energy. The change in the internal energy of liquid and solid bodies (as well as of polyatomic gases) as a result of heating is due to an increase in both the kinetic and the potential energy of these bodies.

creases with heating. Thus, a certain (very small, see Ex. IV.13) fraction of the rest mass of the gas is associated with the kinetic energy of thermal motion of its molecules, which is one of the forms of the internal energy.

The theory of relativity widely generalizes this conclusion and proves that the entire rest mass of a body is proportional to its internal energy. The proportionality factor between the rest mass and the internal energy of the body is the same as between the additional mass  $W_k/c^2$  of the body and its kinetic energy  $W_k$ , i.e.  $1/c^2$ . Consequently,

$$m_0 = \frac{W_{\text{int}}}{c^2}, \quad \text{or} \quad W_{\text{int}} = m_0 c^2, \quad (22.7.1)$$

where  $W_{\text{int}}$  is the internal energy of the body, also called its *rest energy* and  $m_0$  is the rest mass of the body.

Using relation (22.6.1), we can now write

$$m = m_0 + \frac{W_k}{c^2} = \frac{W_{\text{int}}}{c^2} + \frac{W_k}{c^2} = \frac{W}{c^2}.$$

Here  $m$  is the mass of the body and  $W$  is its *total energy* equal to the sum of its internal energy (rest energy) and kinetic energy ( $W = W_{\text{int}} + W_k$ ).

We arrive at *Einstein's law: the mass of a body is proportional to its total energy, or conversely, the total energy of a body is proportional to its mass*. Thus, Einstein's law is expressed by the formula

$$m = \frac{W}{c^2}, \quad \text{or} \quad W = mc^2. \quad (22.7.2)$$

Let us use Einstein's law to determine the rest energy (internal energy) of 1 kg of a substance:

$$W_{\text{int}} = 1 \times (3 \times 10^8)^2 \text{ J} = 9 \times 10^{16} \text{ J}.$$

This energy is enormously large: two million kilograms of the most calorific fuel (petroleum) are required for obtaining such an energy.<sup>7</sup>

In all ordinary processes (chemical reactions, mechanical motion of bodies, etc.), the energy transmitted from one body (or system of bodies) to another body (or system of bodies) is negligibly small in comparison with the rest energy of the bodies participating in a process. It does not exceed a few billionths of the rest energy. Therefore, in ordinary processes the total energy of each participating body changes by not more than a few billionths of its value. For this reason, the mass of bodies, which is proportional to the total energy, remains practically unchanged in such

<sup>7</sup> About  $4.6 \times 10^7$  J are liberated upon combustion of 1 kg of petroleum.

processes (with a very high accuracy). This statement is contained in the mass conservation law which had been discovered by Lomonosov and Lavoisier long before the theory of relativity was created.

In last decades, physicists and engineers have come across phenomena in which energy liberation is so large that it constitutes a noticeable fraction of the rest energy of interacting bodies (atomic energy may serve as an example). In these phenomena, the change in the mass of bodies, which accompanies energy conversion, is also large and can be measured exactly. It will be shown in Secs. 24.6 and 24.8 that the validity of Einstein's law was proved by using such measurements. This law turns out to be very helpful for processes accompanied by the liberation of large amounts of heat. Using Einstein's law, the problem of determining the energy contained in a body is replaced by a much simpler problem of the accurate measurement of its mass. Using (22.6.2), we can write Einstein's law in a different form:

$$W = \frac{m_0 c^2}{\sqrt{1 - v^2/c^2}}. \quad (22.7.3)$$

Let us now establish the relation between the total energy of a body, its rest mass and momentum. Using (22.6.4) and (22.7.3), we shall find the ratio between the velocity of the body and the speed of light:  $W = pc^2/v$ , or  $v/c = pc/W$ . Substituting this expression for  $v/c$  into (22.7.3) for the total energy, we finally obtain a very important relation of relativistic mechanics:

$$W^2 = p^2 c^2 + m_0^2 c^4. \quad (22.7.4)$$

Einstein's law is valid for all objects, i.e. not only bodies or particles, but also, for example, for electric and magnetic fields. According to this law, electromagnetic fields possess a mass. Let us consider, for example, light quanta, viz. the clusters of an electromagnetic wave field. Each quantum of light of a frequency  $\nu$  has an energy  $h\nu$ , where  $h$  is Planck's constant. According to (22.7.2), the quantum  $h\nu$  has a mass  $h\nu/c^2$ . This result was confirmed in experiments.

Light quanta have an important peculiarity: the rest mass of a light quantum is zero. This can easily be verified with the help of the mass-velocity (22.6.2). According to this relation, the rest mass  $m_0 = m\sqrt{1 - v^2/c^2}$ . Light quanta move at the speed of light, i.e.  $v/c = 1$ ,  $\sqrt{1 - v^2/c^2} = 0$  for them, and hence  $m_0 = 0$ .<sup>8</sup>

---

<sup>8</sup> In modern scientific literature, the rest mass  $m_0$  of a body is called just the mass, and the concept of "mass depending on velocity", i.e. (22.6.2), is not used. Instead, the concept of the total energy (22.7.3) is employed. For this reason, we shall henceforth, as a rule, denote the rest mass just by  $m$ .

## 22.8. Mass of Atoms. Isotopes

Let us consider the results of experiments on measuring the masses of positive ions. Figure 352 shows a mass spectrogram of positive neon ions. Three lines of different intensities are clearly seen on the spectrogram. Comparing the distances between the lines and the slit, we can find that lines *A*, *B* and *C* correspond to the values of  $m/e$  which are in the ratio 20:21:22.

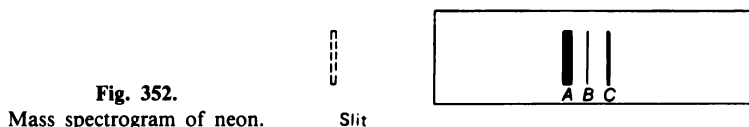


Fig. 352.

Mass spectrogram of neon.

Slit

The presence of three lines cannot be explained by the difference in the charges of the ions. A neon ion can carry a charge which does not exceed a few elementary units.<sup>9</sup> The charge ratio may be 3:2:1 but never  $1/20:1/21:1/22 \approx 22:21:20$ . We have to assume that lines *A*, *B* and *C* are due to the ions carrying the same charge but having different masses with the mass ratio 20:21:22. The atomic mass of neon is 20.2. Consequently, the mean value of the mass of a neon atom is 20.2 amu. But the masses of the ions responsible for lines *A*, *B* and *C* are 20, 21 and 22 amu. We arrive at the conclusion that neon is a mixture of atoms of three types which differ in mass.<sup>10</sup> Comparing the intensities of blackening of lines on the mass spectrogram, we can find the relative amounts of various atoms in natural neon. The numbers of neon atoms with masses of 20, 21 and 22 are in the ratio 90:0.3:9.7.

Let us calculate the mean mass of a neon atom:

$$m_m = \frac{20 \times 90 + 21 \times 0.3 + 22 \times 9.7}{90 + 0.3 + 9.7} = 20.2 \text{ amu.}$$

The fact that  $m_m$  coincides with the atomic mass of neon, which was determined experimentally, confirms the conclusion that neon is a mixture of three types of atoms. It is important to note that the proportion of atoms with masses of 20, 21 and 22 amu is the same for neon samples of different origin (atmospheric neon, neon contained in minerals, and so on). This proportion does not change (or changes insignificantly) in ordinary physical and chemical processes like liquefaction, evaporation or diffusion. This proves that the three modifications of neon are almost identical in their properties.

<sup>9</sup> It will be shown in Sec. 22.15 that a neon atom (having an atomic number of 10 in the Periodic Table) contains only 10 electrons. However, during the gas discharge occurring in the ion source of a mass spectrograph, one electron is detached from a neon atom in most cases and two electrons, in rare cases.

<sup>10</sup> Since the electron mass is very small, the mass of a neutral atom of neon is practically equal to the mass of a positive ion of neon.

Atoms of the same element that differ only in mass are known as *isotopes*. All isotopes of the same element have identical chemical and very close physical properties.<sup>11</sup>

The existence of isotopes is typical not only of neon. Most of elements are mixtures of two or more isotopes. Examples of isotopic composition are given in Table 11.

**Table 11.** Isotopic Composition of Some Elements

| Element  | Atomic mass<br>(rounded) | Isotope                |            |
|----------|--------------------------|------------------------|------------|
|          |                          | mass (rounded),<br>amu | content, % |
| Chlorine | 35.5                     | 35                     | 75.5       |
|          |                          | 37                     | 24.5       |
| Hydrogen | 1                        | 1                      | 99.985     |
|          |                          | 2                      | 0.015      |
| Oxygen   | 16                       | 16                     | 99.76      |
|          |                          | 17                     | 0.04       |
|          |                          | 18                     | 0.20       |
| Uranium  | 238                      | 234                    | 0.006      |
|          |                          | 235                    | 0.720      |
|          |                          | 238                    | 99.274     |

It follows from Table 11 that *the masses of isotopes of all elements are expressed by integral numbers of atomic mass units*. The meaning of this important regularity will be clarified in Sec. 24.8. Accurate measurements indicate that the law concerning the integral numbers of the masses of isotopes is of approximate nature. As a rule, the masses of isotopes slightly differ from integers (in the second to fourth decimal places). In some problems, these small deviations from integers play a decisive role (see, for example, Sec. 24.9).

Nevertheless, in many problems it is possible to use the value of mass rounded off to the closest integral number of atomic mass units. The mass of an isotope in amu (atomic mass) rounded off to an integer is known as the *mass number*.

We have mentioned above that the isotopic composition of neon is nearly constant and that most properties of its isotopes are almost identical. These statements are also valid for all other elements having isotopes.

Isotopes are designated by the chemical symbol of the corresponding element with a superscript indicating the mass number of an isotope. For example,  $^{17}\text{O}$  denotes an oxygen isotope with a mass number of 17,  $^{37}\text{Cl}$

<sup>11</sup> In Chaps. 23 and 24, we shall consider some physical phenomena in which the properties of isotopes of the same element may differ considerably.

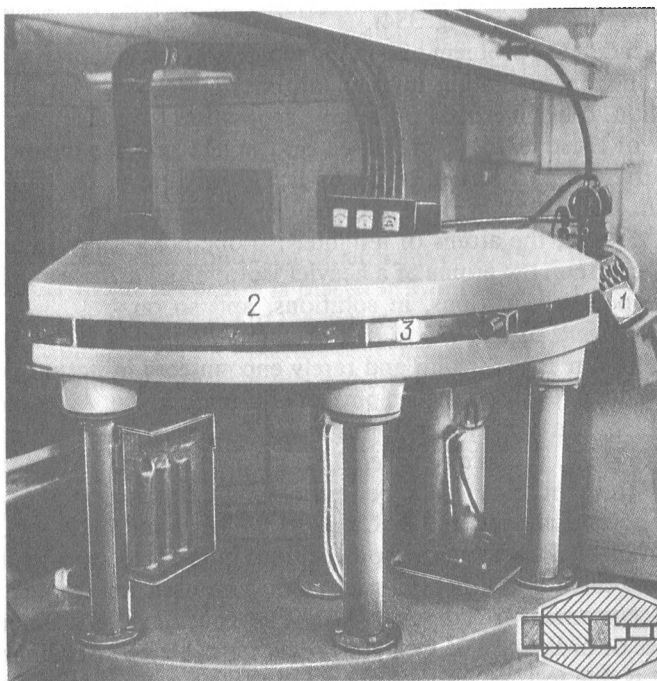
stands for a chlorine isotope with a mass number of 37. Sometimes, the atomic number of the element in the Periodic Table is used as the subscript:  $^{16}_8\text{O}$ ,  $^{17}_8\text{O}$ ,  $^{37}_{17}\text{Cl}$ , etc.

## 22.9. Isotope Separation.

### Heavy Water

All isotopes of a given element enter the same chemical reactions and form chemical compounds which are almost identical in solubility, volatility and other properties used in chemistry for separating elements. For this reason, conventional chemical methods of separation, which are based on the difference in the behaviour of substances in chemical reactions, are inapplicable for separating isotopes of the same element. Isotope separation is therefore a much more difficult problem than the separation of elements.

We have already mentioned one of the methods of isotope separation with the help of a mass spectrograph in which each isotope is represented



**Fig. 353.**

Photograph of one of the first devices for the electromagnetic isotope separation (yielding a few milligrams per day): 1 — ion source, 2 — electromagnet, 3 — vacuum chamber in which the ions make a quarter of revolution in a circle. The cross section of the electromagnet is given in the lower right corner.



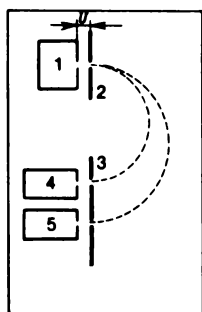


Fig. 354.

Schematic diagram of a device for isotope separation: 1 — ion source, 2 and 3 — diaphragms, 4 — collector of the lighter isotope, 5 — collector of the heavier isotope,  $U$  — voltage accelerating isotopes.

on a photographic plate by an individual line. However, the efficiency of the device shown in Fig. 351 is negligibly low. In order to obtain large amounts of separated isotopes, mass spectrographs of a different design and of a much bigger size are used (Fig. 353). Naturally, the collectors in such devices are vessels with slits at the sites of incidence of ions rather than photographic plates (Fig. 354).

In last decades, the problem of isotope separation has acquired a significant importance in generating nuclear (atomic) power (see Sec. 24.11). In this connection, other methods of isotope separation have also been developed. Most of these methods are based on the fact that the mean kinetic energy of different particles in a gas or liquid mixture is the same, and hence the smaller the mass of a particle, the higher (on the average) its velocity. As a result, the atoms of a lighter isotope possess on the average a higher velocity than the atoms of a heavier isotope and diffuse at a higher rate through porous partitions, in solutions, and so on.

*Heavy hydrogen*, or *deuterium* (the chemical symbol is  $^2\text{H}$  or  $\text{D}$ ), viz. an isotope with a mass of 2 amu and rarely encountered in nature, has acquired a considerable importance in physics and engineering. After combining with oxygen, heavy hydrogen forms water  $\text{D}_2\text{O}$  with a molecular mass of  $2 \times 2 + 16 = 20$ , viz. *heavy water*. This substance noticeably differs in its properties from ordinary water. For example, the freezing point of heavy water under normal pressure is  $3.8^\circ\text{C}$  and its boiling point is  $101.4^\circ\text{C}$ . The biological processes in heavy water proceed not in the same way as in normal water. For this reason, heavy water cannot be used for feeding living organisms accustomed to ordinary water. The considerable difference in the properties of ordinary and heavy hydrogen, and hence of ordinary and heavy water, is due to the fact that the mass of a deuterium atom is twice as large as the mass of a hydrogen atom, while for other elements the mass of a heavy isotope exceeds the mass of a light isotope only insignificantly (for example, only by 5 or 10% for neon).

In electrolysis, heavy water decomposes at a slower rate than ordinary water. This phenomenon is used in a method of obtaining heavy water.

The separation of heavy water is a rather difficult problem since the relative content of heavy water in ordinary water is negligibly low (about a hundredth of percent).

### 22.10. Nuclear Model of Atom

In previous sections, we considered the mass and size of atoms. Let us now discuss the internal structure of atoms. The studies of atomic structure were stimulated by the discovery of radioactivity. This phenomenon will be considered in detail in Chap. 23. So far, we shall confine ourselves to the following facts concerning radioactivity.

Some elements at the end of the Periodic Table are able to emit fast charged particles known as *alpha-particles* ( $\alpha$ -particles). Experiments showed that  $\alpha$ -particles are ionized helium atoms. They carry a positive charge equal to  $2e$  and have a mass of 4 amu. Alpha-particles can be detected by their various actions, for example, by the effect produced by them on luminescent screens. When even one  $\alpha$ -particle strikes a screen coated with a luminescent material (say, zinc sulphide), a short flash called *scintillation* can be seen. Scintillations can easily be detected by eye, especially when observed through a microscope with a small magnification. Alpha-particles are emitted by radioactive atoms at a velocity exceeding 10 000 km/s. Owing to their enormous velocity,  $\alpha$ -particles may penetrate atoms with which they collide. This effect can be used for obtaining information about the internal structure of atoms.

Let us consider the following experiment (Fig. 355). Diaphragm 2 with a small aperture at the centre is placed in front of source 1 of  $\alpha$ -particles. The alpha-particles arriving to the material of the diaphragm are captured, while those incident on the aperture pass through it in the form of a narrow beam. A luminous spot caused by the scintillations emerging under the impacts of individual  $\alpha$ -particles is formed in the region where the beam of  $\alpha$ -particles is incident on a transparent luminescent screen 3. Since the number of particles incident on the screen per second is large, individual scintillations merge into a luminous spot.

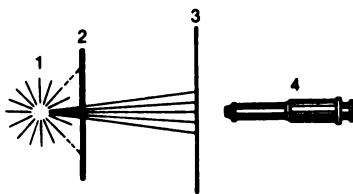


Fig. 355.

Observation of scintillations caused by  $\alpha$ -particles: 1 — source of  $\alpha$ -particles, 2 — diaphragm with a small aperture, 3 — luminescent screen, 4 — microscope for observing scintillations.

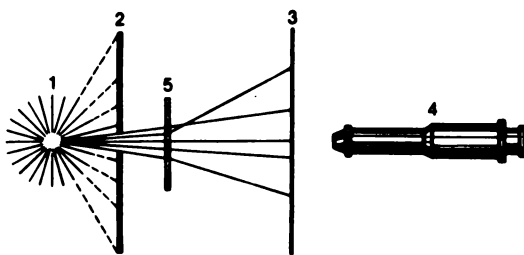


Fig. 356.

Scattering of  $\alpha$ -particles by gold foil 5 (other notations are the same as in Fig. 355).

Let us now place in front of the screen a thin layer of a substance, for example, a gold foil having a thickness of about  $1\text{ }\mu\text{k}$ . It will be seen (Fig. 356) that the intensity of the central luminous spot will decrease although insignificantly. At the same time, some scintillations will be observed outside the central spot. These scintillations are due to  $\alpha$ -particles that have changed their direction as a result of the passage through the gold foil (or have been scattered). By moving the microscope over the screen from the central spot outwards, we find that the number of scattered  $\alpha$ -particles rapidly decreases with increasing scattering angle.

In the above experiment, the following is remarkable. The diameter of a gold atom is  $3 \times 10^{-10}\text{ m}$ . The gold foil  $1\text{ }\mu\text{m}$  thick contains  $10^{-6} \div (3 \times 10^{-10}) = 3300$  atomic layers. Atoms in solids are arranged almost closely (see Sec. 22.2). Therefore, an  $\alpha$ -particle passing through the foil must collide with about 3000 atoms of gold. Nevertheless, it was shown that the major fraction of  $\alpha$ -particles pass through the foil without undergoing a noticeable scattering. The results of this experiment allow us to conclude that an atom of gold can by no means be regarded as impenetrable.

On the other hand, it is important to note that some  $\alpha$ -particles passing through the foil are scattered at large angles. In order to deflect an  $\alpha$ -particle passing at a huge velocity through a large angle, enormous forces are required. Consequently, within an atom an  $\alpha$ -particle may experience the action of huge forces, but only a small fraction of passing particles get into the field of these forces.

In order to explain these experiments, the English physicist Sir Ernest Rutherford (1871-1937) proposed a *nuclear model of atomic structure* in 1911. According to this model, almost the entire mass of an atom is concentrated in its positively charged nucleus that occupies a negligible part of the volume of the atom. The positive nucleus is surrounded by negative electrons. The electron shell occupies virtually the entire volume of the atom, but the mass of the shell is insignificant because of the small mass of an electron.

**Fig. 357.**

Collision between a heavy and a light ball. Solid arrows indicate the velocities of the balls before the collision (a) and after the collision (b). The motion of the heavy ball changes only slightly as a result of collision (dashed arrows indicate the velocities of the balls before the collision).

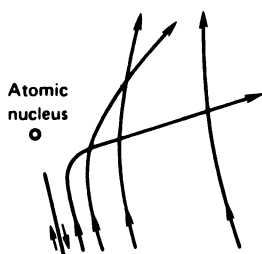
Let us consider the passage of an  $\alpha$ -particle through an atom from the point of view of such a nuclear model. The  $\alpha$ -particle penetrating the atom is acted upon by the electric forces exerted by the nucleus and electrons.<sup>12</sup> The mass of an electron is equal to about  $1/8000$  of the mass of an  $\alpha$ -particle. Therefore, the interaction between the  $\alpha$ -particle and the electron is similar to the elastic collision of a rapidly moving heavy ball with a light ball. In such a collision, the direction of motion of the light ball may change drastically, while the velocity of the heavy ball changes insignificantly (Fig. 357). Thus, the interaction with the electrons does not lead to a noticeable deflection of the  $\alpha$ -particle. As regards the interaction between the  $\alpha$ -particle and the nucleus, it may noticeably change the motion of the  $\alpha$ -particle. Indeed, in the case of gold, the role of the heavy ball is played by the nucleus of the gold atom, while the role of the light ball is played by the  $\alpha$ -particle (the mass of a gold atom is 197 amu, while the mass of an  $\alpha$ -particle is 4 amu).

The deflection of the  $\alpha$ -particle is proportional to the force acting on it, which is the stronger, the closer the  $\alpha$ -particle to the nucleus.

The fact that some  $\alpha$ -particles undergo considerable deflections proves that sometimes an  $\alpha$ -particle and a nucleus may approach each other to a very short distance, i.e. the sizes of both  $\alpha$ -particles and the nucleus are very small. However, the number of  $\alpha$ -particles passing close to the nuclei is small. Most  $\alpha$ -particles pass at a comparatively large distance from the nuclei, and hence are deflected insignificantly (Fig. 358).

Using Coulomb's law and Newton's laws of dynamics, Rutherford calculated the number of scattered  $\alpha$ -particles as a function of the scattering angle. The results of calculations are in excellent agreement with the results of measurements with the foils of different materials. This coincidence proves the correctness of the nuclear model of atom. It also confirms the

<sup>12</sup> Gravitational forces also act between particles within an atom, but they are so weak in comparison with the electric forces that can be neglected in the case under consideration (see Exercise IV.18).



**Fig. 358.**  
Trajectories of  $\alpha$ -particles passing by  
an atomic nucleus at different dis-  
tances.

correctness of the assumption that the electric forces acting within an atom obey Coulomb's law ( $\sim 1/r^2$ ). It is well known, however, that Coulomb's law is valid only when the sizes of interacting charges are small in comparison with the distances separating them. The fact that this law is satisfied even when the centres of the interacting nucleus and  $\alpha$ -particle are separated by very short distances indicates that the size of the nucleus must be very small. Theoretical calculations and their comparison with experimental results allow us to draw quantitative conclusions about the size of a nucleus and its charge.

It turns out that the diameters of nuclei of different atoms have slightly different values (the diameter of a nucleus is the larger, the larger the atomic mass) and are of the order of  $10^{-12}$  cm. Therefore, the size of a nucleus is about 1/10 000 the size of the atom. Let us imagine for a moment that we can look inside a dense medium, viz. a liquid or a solid. We shall see a "mist" of light electrons filling the entire volume of the substance. In this "mist", we shall see sparsely located very small but heavy atomic nuclei separated from one another by distances exceeding 10 000 times their size.

The charge of a nucleus is  $+Ze$ , where  $e$  is the elementary charge and  $Z$  is the atomic number of the element in the Periodic Table. Since the atom is neutral as a whole, the number of electrons in it is  $Z$ . Thus, the atomic number of an element in the Periodic Table has a profound physical meaning: *the atomic number of an element is the charge of its nucleus in elementary units of charge and at the same time the number of electrons in the atom.*

### 22.11. Energy Levels of Atoms

Experiments on the scattering of  $\alpha$ -particles revealed the presence of a heavy positive nucleus and an electron shell in an atom. A further information about the properties of atoms was obtained from the analysis of atomic processes accompanied by a change in the internal energy of an atom. They include collisions between atoms and electrons, emission and absorption of light by atoms, and so on. Investigating these processes, scientists succeeded in establishing peculiar and very important regularities governing the internal energy of atoms.

**Collisions between electrons and atoms.** The analysis of energy transfer from electrons to atoms can easily be observed with the help of a device shown in Fig. 359. A small amount of monatomic vapour of some substance, say, mercury,<sup>13</sup> is introduced into tube 1 from which air has been pumped out. The electrons emitted by incandescent cathode 2 are accelerated by the potential difference  $U_1$  created between cathode 2 and metal grid 4. Due to a small concentration of atoms, electrons cover a short distance between the cathode and the first grid almost without collisions and acquire energy  $eU_1$ .

The electric field behind the first grid 4 and in the space between it and the second grid 5 is zero since the grids are under the same potential, and the energy of an electron may change only as a result of its collision with an atom. The distance between the grids is chosen sufficiently long so that every electron undergoes at least one collision. Further, a potential difference  $U_2$  is created between the second grid and the anode, which decelerates electrons. As a result, only the electrons having an energy exceeding  $eU_2$  can reach the anode.

By gradually increasing  $U_2$ , we shall determine the cutoff potential difference, i.e. the minimum value of  $U_2$  for which no electrons reach the anode and the current through the galvanometer ceases. By measuring the cutoff potential, we can establish whether or not electrons lose their energy as a result of collisions with atoms. Indeed, if the electrons do not lose energy on their way between the grids, the cutoff potential will be equal to the accelerating potential. Otherwise, it will be smaller. If in this case every electron gives away energy  $W$ , the difference between the accelerating and decelerating potentials will be  $W/e$ .

Experiments of this kind with mercury vapour led to a remarkable result. It turned out that the energy transfer from electrons to atoms con-

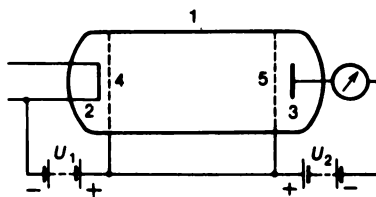


Fig. 359.

Device for measuring the energy lost by an electron moving in mercury vapour: 1 — glass tube filled with mercury vapour (under a pressure of the order of  $10^{-3}$  mmHg), 2 — incandescent cathode (heater is not shown in the figure), 3 — anode, 4 and 5 — sparse metal grids connected to each other,  $U_1$  and  $U_2$  are accelerating and decelerating potentials.

<sup>13</sup> Experiments can be carried out not only with atoms but with molecules as well. In this case, however, phenomena are considerably more complicated. For this reason, we shall confine ourselves to the case of monatomic substances.

siderably depends on the electron energy. As long as electrons have an energy lower than 4.9 eV (i.e.  $U_1 < 4.9$  V), they do not lose energy at all as a result of collisions with atoms (i.e.  $U_2 = U_1$ ). When, however, the energy of the electrons attains (or slightly exceeds) 4.9 eV (i.e.  $U_1 \geq 4.9$  V), the energy loss in collisions becomes significant (i.e.  $U_2 \ll U_1$ ). In this process, an electron gives away, and hence a mercury atom receives, always the same amount of energy equal to 4.9 eV<sup>14</sup>. This quantity obviously characterizes a property of mercury atoms. Their energy may change only by a finite value equal to 4.9 eV. Mercury atoms do not accept a lower energy.

While studying mechanics, heat and electricity, we have not come across such a phenomenon: the energy of any body or a system of bodies could, in principle, change continuously, i.e. by as small portions as desired. But in the case of a mercury atom, a continuous change in energy is impossible since the energy of the mercury atom changes discretely, i.e. by a finite value.<sup>15</sup>

Carrying out similar experiments with other substances, we arrive at the same conclusion about the discrete nature of the energy levels of atoms.

**Analysis of optical spectra.** It is well known (see Sec. 20.2) that elements in the gaseous state have line emission and absorption spectra. Every element is characterized by definite spectral lines differing from the lines of other elements. Since the atoms of a gas are on the average at large distances and do not affect one another, the frequencies of the line spectrum of an element must be determined by the properties of an individual atom of the element.

It was shown in Chap. 21 that the luminous energy exists in the form of very small indivisible portions, viz. quanta. Consequently, atoms must emit and absorb light in the same portions (quanta). The energy of a quantum is proportional to the frequency  $\nu$  of light, i.e. is equal to  $h\nu$  where  $h \approx 6.6 \times 10^{-34}$  J·s is Planck's constant. According to the energy conservation law, the energy of a quantum emitted by an atom is equal to the difference in the energies of the atom before and after the act of emission:

$$h\nu = W - W', \quad (22.11.1)$$

where  $W$  is the energy of the initial state of the atom (before the emission) and  $W'$  is the energy of the final state of the atom (after the emission).

Relation (22.11.1) connects the change in the energy of an atom as a result of emission or absorption of light with the light frequency  $\nu$ . If the

<sup>14</sup> If the energy of electrons exceeds 6.7, 8.3, etc. electron volts, the portions of energy transferred as a result of collisions with mercury atoms may be not only 4.9 but also 6.7, 8.3, etc. eV.

<sup>15</sup> Here we speak of the internal energy of an atom. The kinetic energy of an atom moving as a single whole may vary as little as desired because the velocity of translational motion of the atom may change by any small value.

energy of the atom could change arbitrarily, various frequencies would be present in the atomic spectrum, and the latter would be continuous as the spectrum of incandescent solid body. In actual practice, however, the atomic spectrum (i.e. the emission or absorption spectrum of a monatomic gas) is a line spectrum rather than a continuous one. It contains only certain frequencies characteristic of a given atom. Consequently, the energy of an atom cannot be changed arbitrarily. It can rather be changed by definite values. If we know the spectrum of a substance, we can easily determine these values with the help of (22.11.1).

For example, the absorption spectrum of mercury vapour contains the following lines (in the order of decreasing wavelength): 253.7, 185.0, 140.3 nm, etc. Substituting the values of frequency into (22.11.1), we obtain for the first line

$$W - W' = h\nu = \frac{hc}{\lambda} = \frac{6.6 \times 10^{-34} \times 3 \times 10^8}{253.7 \times 10^{-9}} = 7.8 \times 10^{-19} \text{ J},$$

or

$$W - W' = \frac{7.8 \times 10^{-19}}{1.6 \times 10^{-19}} = 4.9 \text{ eV}.$$

For the second and third lines, we obtain  $W - W' = 6.7 \text{ eV}$  and  $W - W' = 8.3 \text{ eV}$  respectively. Thus, a mercury atom can accept energy only in portions of 4.9, 6.7, 8.3, . . . , eV. The minimum acceptable portion turns out to be 4.9 eV in accordance with the results obtained in experiments on collisions between electrons and atoms.<sup>16</sup>

Thus, the two classes of phenomena considered by us, viz. optical spectra and interaction between atoms and electrons, point to the discrete nature of the internal energy of atoms. The energy of an atom cannot change continuously. It changes abruptly by certain finite portions which are different for different atoms. It follows hence that the energy of an atom cannot be arbitrary, but rather assumes certain selected values, typical of every atom. The possible (allowed) values of the internal energy of an atom are termed *energy*, or *quantum*, *levels* (*states*).

The energy level diagram for a hydrogen atom, constructed on the basis of spectral lines, is shown in Fig. 360 in the form of a series of parallel lines. The separation between two lines is equal to the difference in the energies corresponding to two states of the hydrogen atom, and hence is proportional to the frequency of a quantum emitted upon a transition from one state to another (lower) state. For this reason, the separations between

<sup>16</sup> It should be noted that in these experiments, mercury vapour glows, emitting light of a wavelength of 253.7 nm, which appears when the energy of electrons becomes equal to or higher than 4.9 eV.



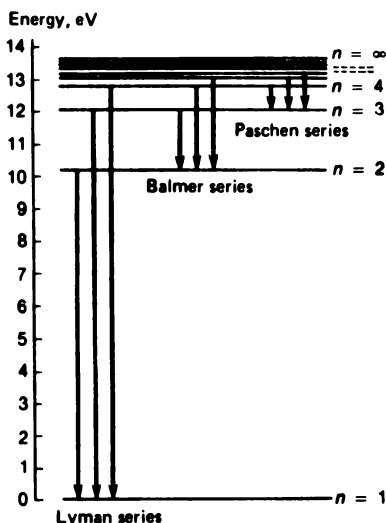


Fig. 360.

Energy level diagram for a hydrogen atom. Horizontal lines indicate energy levels ( $n$  is the level number). The origin on the energy scale corresponds to the minimum internal energy of the hydrogen atom, i.e. the energy of a level with  $n = 1$ . Vertical lines indicate the transitions from the upper to the lower energy levels. The length of such a segment gives the energy  $h\nu$  of a light quantum emitted in a given transition. The transitions are grouped into series: the Lyman series (transitions from levels with  $n > 1$  to the ground level with  $n = 1$ ), the Balmer series (transitions from levels with  $n > 2$  to the level with  $n = 2$ ), and so on (see also Sec. 20.4).

levels are plotted on a certain scale of frequency of hydrogen spectral lines. After a short time (of about  $10^{-8}$  s) an atom in one of the higher energy states (denoted by a number  $n > 1$  in Fig. 360) goes over to a lower energy state, emitting thereby a relevant quantum. From the lowest energy state ( $n = 1$ ), the atom cannot go over to another state spontaneously (without receiving an energy from outside). Consequently, the lowest state is stable.<sup>17</sup> Under normal conditions, all atoms are in the lowest energy state, and the gas does not glow.

By imparting an energy to an atom, we can excite it, i.e. transfer from the normal (lowest) state to one of higher energy states. In the case of hydrogen, the separation between the lowest energy state ( $n = 1$ ) and the nearest (higher) energy state is 10.1 eV. This is the smallest portion of energy that a hydrogen atom can absorb in the lowest energy state. This atom cannot accept a smaller amount of energy since it does not have states with energies differing from the energy of the normal state by less than 10.1 eV. As was shown above, the relevant value for a mercury atom is 4.9 eV.

<sup>17</sup> The lowest energy state is known as the *ground state*, while all the remaining states are termed *excited states*.

## 22.12. Induced Emission of Light.

### Quantum Generators

The idea about quantum energy levels was introduced in 1913 by Niels Bohr. He provided a natural explanation of line atomic spectra as a result of *spontaneous emission* and *resonance (selective) absorption* of light by atoms (Figs. 361a and b). In 1919, Einstein proved that along with spontaneous emission and resonance absorption, there also exists a third process, viz. *induced emission*. According to Einstein, the light of a resonance frequency, i.e. the frequency that can be absorbed by atoms (which, as a result, go to a higher energy level), must cause emission of photons by the atoms occupying this upper level (Fig. 361c) if such atoms are present in the medium.

A typical feature of induced emission is that *the emitted light cannot be distinguished from the light causing emission*, i.e. it coincides with it in all properties, viz. frequency, phase, polarization and the direction of propagation. This means that the induced radiation supplies to the luminous flux the same quanta as those carried away from it by the resonance absorption. For this reason, only the difference between the absorbed and induced radiation is manifested in experiments. Atoms occupying the lower of the two levels participating in the process absorb light, while atoms occupying the upper level emit light. Therefore, *the absorption prevails, and the light beam is absorbed by a medium if it contains more atoms on the lower level than on the upper level*.

On the contrary, *if the upper level is populated more densely, induced emission prevails, and the medium amplifies the light passing through it*.

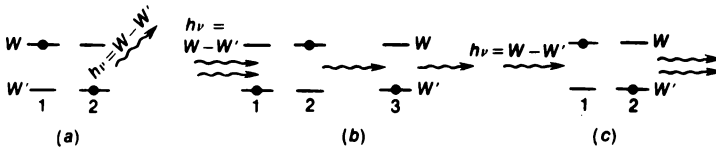


Fig. 361.

(a) Spontaneous emission of light: 1 — as a result of the collision with another atom or with an electron, or as a result of the absorption of a light quantum, an atom goes to one of its upper levels  $W$  (for the sake of simplicity, only two quantum levels  $W$  and  $W'$  are shown in the figure), 2 — in a certain time, the excited atom goes by itself (spontaneously, without external effects) to the lower level  $W'$ , emitting a light quantum  $h\nu = W - W'$ . (b) Resonance absorption of light: 1 — an atom occupying level  $W'$  is irradiated by light of frequency  $\nu = (W - W')/h$  (resonance frequency), 2 — the atom absorbs a light quantum  $h\nu = W - W'$  from the light beam and goes to level  $W$ ; the light beam is attenuated, 3 — the excited atom collides with another atom, transfers its energy to it and returns to level  $W'$  (it may return to  $W'$  or go to another level lying below  $W$ ; also emitting a light quantum). (c) Induced emission of light: 1 — an atom occupying the upper level  $W$  is irradiated by light of the resonance frequency, 2 — the atom emits a quantum  $h\nu = W - W'$  in the direction of the incident light beam which is thus augmented; the atom goes thereby to the lower level  $W'$ .

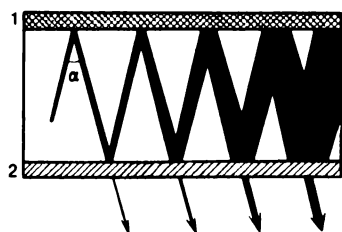


Fig. 362.

Schematic diagram of laser operation: 1, 2 — plane-parallel mirrors; mirror 2 is semitransparent. (Angle  $\alpha$  in the figure is made much larger than in reality.)

Recently, this phenomenon has been used in very promising devices known as *quantum generators of light (lasers)*<sup>18</sup>. The schematic diagram of operation of a laser is shown in Fig. 362. The space between mirrors is filled with an *active medium* containing a larger number of excited atoms than unexcited atoms. The medium amplifies the light emitted spontaneously by one of the atoms. Strong amplification is attained when angle  $\alpha$  is very small so that light undergoes multiple reflections, and the rays are superimposed, augmenting one another. (This corresponds to the formation of a standing light wave in the space between plates 1 and 2, see Sec. 5.5.) Laser radiation emerges through mirror 2. Such a laser emits light at a frequency  $\nu = (W - W')/h$ , where  $W - W'$  is the difference in the energies of the levels participating in the process. Lasers operating in the range of radio waves, infrared, visible and ultraviolet radiation have been constructed.

Since atoms go from the upper to the lower level as a result of emission, this leads to a rapid depletion of the upper level which previously had an excess population. If this depletion is not compensated, the laser stops emitting as soon as the excess population is reduced to a certain limiting value.

The principles formulated above were realized and implemented only after 30-40 years following the discovery of induced emission by Einstein.<sup>19</sup> The reason behind such a time lag is the abnormality of the state in which most of atoms are on the upper energy level. Under normal conditions, the lower level is populated more densely. This is due to the fact that a portion of energy equal to the energy difference  $W - W'$  between the levels has to be supplied so that an atom can go from the lower to the upper level, while the inverse transition does not require any energy to be supplied. At low temperatures, only an insignificant fraction of atoms possess a kinetic energy higher than  $W - W'$ . For this reason, the excitation of an atom as a result of atomic collisions is an exceptionally rare event, and all the atoms are on the ground level. This is manifested in the well-known

<sup>18</sup> These devices are known as *lasers (masers)* which is an abbreviation from "light (micro-wave) amplification by stimulated emission of radiation".

<sup>19</sup> This was stimulated by the research work carried out by several groups of scientists in the USSR (N. G. Basov, A. M. Prokhorov and V. A. Fabrikant), the USA and Canada.

fact that cold bodies do not glow. As the temperature rises, the population of excited levels increases, and the body starts to glow.

At very high temperatures, when the kinetic energy of atoms is much higher than  $W - W'$ , the energy required for exciting an atom as a result of atomic collisions can easily be attained. Under such conditions, the populations of levels are equalized. However, it is still impossible to populate the upper level more (densely) than the lower level by heating. This can be done only by using special methods one of which will be considered here while describing a ruby laser.

Ruby is a crystal of aluminium oxide  $\text{Al}_2\text{O}_3$  (corundum) containing small quantities of chromium oxide. The operation of a ruby laser is based on the transition between energy levels of chromium atoms (to be more precise, ions) in the crystal. The energy level diagram for chromium is shown in Fig. 363. By exposing the crystal to green light, one can transfer a chromium atom from the ground level 1 to level 3. In most cases, the atom goes over from level 3 to level 2 through a nonradiative transition, giving away its energy to the corundum lattice, and then from level 2 to level 1. The rate of transition  $3 \rightarrow 2$  is thousands of times higher than the rate of transition  $2 \rightarrow 1$ . As a result, atoms are accumulated on level 2. If the crystal is exposed to green and blue light of a very high intensity, more than half of chromium atoms contained in the crystal can be transferred to level 2. In other words, we can create a population inversion of the levels which is required for lasing. The hydraulic analogy of the method

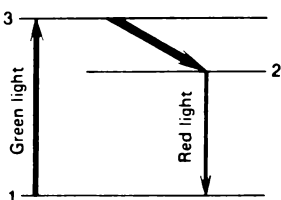


Fig. 363.

Energy level diagram of a chromium atom in ruby.

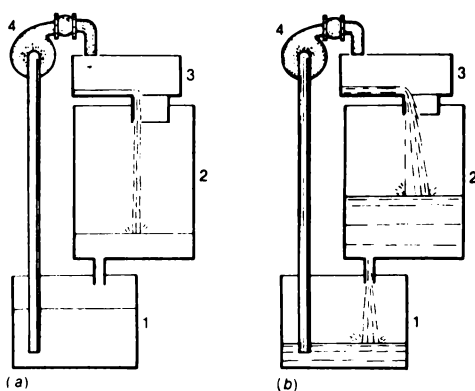


Fig. 364.

The hydraulic analogy of the method of optical "pumping" for creating an excess population of the excited level. The height of water column in tank 2 increases during the operation of pump 4 until it becomes high enough to "force" all water delivered by the pump through a narrow discharge pipe. Operation of the pump with a low (a) and a high (b) capacity. In this model, tanks 1, 2 and 3 play the role of energy levels of a chromium atom in ruby, and the heights of the water columns, the population of the levels. The pump can be treated as a source of "pumping" green light.

of optical “pumping” for obtaining an excess population of the excited level is shown in Fig. 364.

The ruby laser is shown schematically in Fig. 365. A “pumping” pulse of green and blue light emerges when capacitor 1 is discharged through a pulsed gas-discharge tube (flash lamp) 2 placed in a reflecting jacket 3. Tube 2 has the shape of a spiral wound on a ruby rod 4 having strictly plane-parallel polished end faces with mirror layers deposited on them. As soon as a sufficiently large excess of atoms is accumulated on level 2 over level 1 under the action of a “pumping” flash (see Fig. 363), the above process of light emission at a frequency corresponding to the difference between the energy levels 2 and 1 (viz. red light with a wavelength of 690 nm) is observed. A narrow red light beam will be emitted by the ruby rod through one of its end faces (whose mirror coating is semitransparent). The beam is parallel to a high degree since generation occurs due to multiple reflection of the waves passing along the crystal and being reflected at the mirrors at its ends, i.e. propagating perpendicularly to the end faces of the ruby rod (see Fig. 365). Naturally, the emission from such a laser will last only during the discharge of capacitor 1 through the gas-discharge tube 2. In other words, the laser will operate in a pulse mode. For some other mechanism of excitation, a continuous lasing can be ensured. Most easily, this is attained with the help of gas lasers.

A very important property of laser radiation is its *coherence* (see Secs. 5.2 and 13.2). Light waves emitted by different regions of the luminous surface of a laser are in phase, and oscillations are regular in the sense that their frequency is constant and the phase does not experience abrupt changes. In this respect, lasers are superior to all other light sources and essentially do not differ from ordinary radiowave generators. The coherence of laser radiation is due to the fact that the emitted light is strictly matched with inducing radiation and is indistinguishable from it. The coherence of laser radiation is so high that sometimes the interference of two light beams emitted by independent generators can be observed for some types of lasers. As was pointed out in Sec. 13.2, this can never be done with ordinary sources of light.

Coherence, monochromaticity and collimation of laser radiation make it possible to focus it with the help of converging lenses to a small area

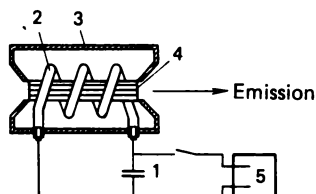


Fig. 365.

Schematic diagram of a ruby laser: 1 — capacitor, 2 — gas-discharge tube, 3 — reflecting jacket, 4 — ruby rod, 5 — energy supply for charging capacitor 1.

of the order of the square of the wavelength. The energy concentration at the focus turns out to be so high that a beam from a ruby laser focussed on a steel plate immediately burns a very small hole in it.

These properties of lasers also determine their other applications such as energy transmission and communication over large distances right up to astronomical range. These prospects stimulate serious efforts of physicists and engineers directed towards further improvement of lasers.

Einstein arrived at the conclusion about the existence of induced radiation as a result of the analysis which can be presented in a simplified form as follows. Let us consider an opaque vessel with two channels 1 and 2, filled with a gas and placed into a thermostat (Fig. 366). Suppose that the same temperature sets in in all parts of the system. In the further process, the temperature in the vessel cannot change spontaneously since some work has to be done to create temperature difference (i.e. to transfer heat from a cold to a hot body) (see Vol. 1, Chap. 19). Since the heated thermostat glows, a radiation is emitted by it. The intensity of the radiation beam entering the vessel through channel 1 must be equal to the intensity of the beam emerging from channel 2. Otherwise, some energy would be introduced into (or carried away from) the vessel, and the temperature in it would change, which is impossible. But the gas atoms in the vessel, which are on the lower energy level, absorb light at the resonance frequency, thus suppressing the radiation emerging from channel 2. Consequently, this absorption must be compensated by radiation. The spontaneous emission of atoms occupying the upper level cannot yield a complete compensation.

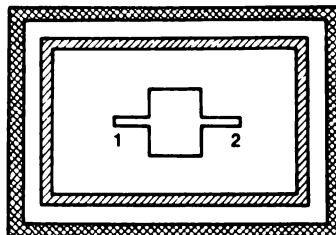


Fig. 366.

To the emergence of induced radiation.

Indeed, with increasing temperature, the intensity of spontaneous emission stops growing as soon as the populations of the upper and lower levels become equal. At the same time, the intensity of thermal glow increases indefinitely with temperature. The intensity of the light beam entering the vessel and the number of quanta absorbed per second increase in the same proportion. For this reason, at a very high temperature spontaneous emission can be disregarded. Hence it follows that there must be a radiation at the same resonance frequency, whose intensity is proportional to the luminous intensity of the beam, which compensates absorption completely when the populations of the upper and lower levels in atoms are equal. This is just the induced radiation.

## 22.13. Hydrogen Atom.

### Peculiarities of Motion of an Electron in an Atom

The existence of discrete energy levels is a fundamental property of atoms (as well as of molecules and atomic nuclei).

Let us try to apply the well-known physical laws to explain the structure of an atom with discrete energy levels.

We shall consider the simplest atom, viz. a hydrogen atom. The atomic number of hydrogen in the Periodic Table is unity. This means that the hydrogen atom consists of a positive nucleus having a charge  $+e$  and an electron. The nucleus and the electron attract each other in accordance with Coulomb's law. The presence of an attractive force ensures a radial (centripetal) acceleration, and as a result, the light electron revolves around the heavy nucleus along a circular or elliptical orbit in the same way as a planet revolves around the Sun under the action of a gravitational force. Therefore, various possible states of the atom correspond to different sizes (and shapes) of the orbit of the electron revolving around the nucleus.<sup>20</sup>

The energy of the electron in an atom is the sum of the kinetic energy of its orbital motion and the potential energy in the electric field of the nucleus. It can be shown (see the end of the section) that the energy of the electron moving in a circular orbit and hence the energy of the atom as a whole, depends on the orbital radius: *a smaller orbital radius corresponds to a lower energy of the atom*. But as was proved in Sec. 22.11, the energy of an atom may assume not arbitrary but only definite selected values. Since the energy is determined by the orbital radius, each energy level of the atom corresponds to an orbit of a strictly definite selected radius.

The pattern of possible circular orbits of the electron in a hydrogen atom is shown in Fig. 367. The ground energy state of the atom corresponds to the orbit of the minimum radius.

Under normal conditions, the electron is in this orbit. If a sufficiently large portion of energy is transferred to it, the electron goes to a higher energy level, i.e. "jumps" to one of the outer orbits. As was mentioned above, the atom in such an excited state is unstable. After a certain time, the electron goes over to a lower level, i.e. "jumps" to an orbit of a smaller radius. A transition of the electron to a closer orbit is accompanied by the emission of a light quantum.

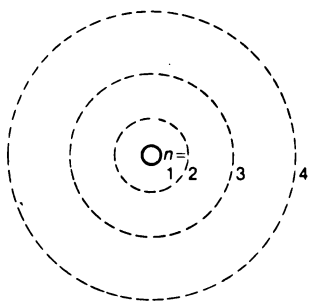


Fig. 367.

Possible orbits of the electron in a hydrogen atom: the orbital radius increases in proportion to  $n^2$ , i.e. in the ratio 1:4:9:16, etc.

<sup>20</sup> In the further discussion, we shall assume that all the orbits are circular. This simplification does not affect the conclusions.

Thus, the nuclear model of the atom and the discreteness of its energy levels imply the existence of selected, "allowed" electron orbits in the atom. But why cannot the electron revolve around the nucleus in an orbit of an arbitrary radius? What is the difference between the allowed and forbidden orbits from the physical point of view?

The laws of mechanics and electricity considered in the previous parts of this book (see Vols. 1 and 2) do not provide any answer to these questions. From the point of view of these laws, all orbits are equivalent. The existence of selected orbits contradicts these laws.

Another drastic contradiction to the well-known laws is the stability of the atom (in the ground state). Any charge moving with an acceleration is known to emit electromagnetic waves. Electromagnetic radiation carries away an energy. The electron in an atom moves at a high velocity in an orbit of a small radius, and hence possesses a huge centripetal acceleration. According to the well-known laws, the electron must lose its energy, emitting it in the form of electromagnetic waves. But as was mentioned above, if the electron loses energy, its orbital radius decreases. Consequently, the electron cannot move in an orbit of a constant radius. Calculations show that as a result of a decrease in the orbital radius, due to emission the electron would fall into the nucleus after  $10^{-8}$  s. This conclusion sharply contradicts our everyday experience concerning the stability of atoms.

Thus, there is a contradiction between the experimental results on the atomic structure and the fundamental laws of mechanics and electricity, which were also derived on the basis of experiments.

It should be borne in mind, however, that the laws mentioned above were derived and verified in experiments with bodies containing a very large number of electrons and atoms.<sup>21</sup> We have no ground to assume that these laws are applicable to the motion of an individual electron in an atom. Moreover, the discrepancy between the behaviour of an electron in an atom and the laws of classical physics<sup>22</sup> points to the fact that these laws are inapplicable to atomic phenomena (see also Sec. 22.17).

At the beginning of the section, we described the *planetary model* of the atom, i.e. the concept of electrons revolving in allowed orbits around an atomic nucleus.<sup>23</sup> This model was substantiated by the laws of classical physics. But as was mentioned earlier and as will be shown in greater detail in Sec. 22.17, the motion of an electron in an atom is one of the phenomena to which classical mechanics is inapplicable. It is not surprising therefore

---

<sup>21</sup> Such bodies are referred to as *macroscopic*.

<sup>22</sup> The laws of classical physics are the laws established for macroscopic bodies. However, quantum phenomena are sometimes (as in the case of liquid helium) exhibited on macroscopic scale also.

<sup>23</sup> The planetary model of the atom was proposed in 1913 by E. Rutherford and N. Bohr.



that a deeper investigation into the “microworld” has revealed the incompleteness and approximate nature of the planetary model. Actually, the atomic structure is more complicated. Nevertheless, the model correctly reflects many of the basic properties of atoms and is sometimes used in spite of its approximate nature.

Let us consider the dependence of the energy of a hydrogen atom on the radius of the electron orbit. The kinetic energy of the motion of the electron in an orbit of radius  $r$  is determined from the condition that the centripetal acceleration  $v^2/r$  is due to the force  $ke^2/r^2$  of the Coulomb attraction of charges (in SI,  $k = 1/4 \pi \epsilon_0$ ). Equating the acceleration  $ke^2/mr^2$  produced by this force to the centripetal acceleration  $v^2/r$ , we find that the kinetic energy  $mv^2/2$  of the electron is inversely proportional to the orbital radius, i.e.  $mv^2/2 = ke^2/2r$ .

Let us select two orbits of radii  $r + a$  and  $r - a$ . The kinetic energy of the electron revolving in the second orbit is higher than the energy of the electron in the first orbit by

$$\frac{ke^2}{2(r-a)} - \frac{ke^2}{2(r+a)} = \frac{kae^2}{r^2 - a^2}.$$

If the orbits are separated by a small distance such that  $a \ll r$  and  $a^2 \ll r^2$ , we can neglect the quantity  $a^2$  in the denominator, and the difference in kinetic energy will approximately be  $kae^2/r^2$ .

On the contrary, the potential energy of the electron is higher for the first, remote orbit since in order to remove the electron from the nucleus, a work has to be done against the forces of electrostatic attraction exerted by the nucleus on the electron. This work is spent on increasing the potential energy.

Let the electron be transferred from the closer orbit to the farther orbit along the radius. The path length is  $(r + a) - (r - a) = 2a$ . The electric force acting along this path is not constant in magnitude. But since the orbits are close to each other ( $a \ll r$ ), we can approximately calculate the work by using the value of the force corresponding to the mean distance between the electron and the nucleus, equal to  $[(r + a) + (r - a)]/2 = r$ . According to Coulomb's law, the force is  $ke^2/r^2$ , and the work done over the distance  $2a$  and equal to the increment of the potential energy is  $k \cdot 2ae^2/r^2$ .

[Thus, as the electron goes over to the closer orbit, the decrease in its potential energy is equal to twice the increment in its kinetic energy.] This theorem was proved by us for close orbits with a separation satisfying the condition  $2a \ll r$ . Summing up the changes in the electron energy for the transitions between consecutive pairs of close orbits, we see that this theorem is also valid for as far orbits as desired.

Let us now consider an infinitely remote orbit, i.e.  $r \rightarrow \infty$ . The potential energy of the electron on this orbit will be taken as the zero potential energy, i.e. we put  $W_p(r = \infty) = 0$ . The kinetic energy  $mv^2/2 = ke^2/2r$  vanishes for  $r = \infty$ . As the electron goes over from the orbit  $r = \infty$  to an orbit having a finite radius  $r$ , its kinetic energy increases by  $ke^2/2r$ . The potential energy decreases thereby by twice as large value ( $ke^2/r$ ), i.e.

$$\begin{aligned} W(r) &= W_p(r = \infty) - k \frac{e^2}{r} = 0 - k \frac{e^2}{r}, \\ W(r) &= -k \frac{e^2}{r}. \end{aligned} \quad (22.13.1)$$

Consequently, the total energy of the electron is

$$\frac{mv^2}{2} + W(r) = k \frac{e^2}{2r} - k \frac{e^2}{r} = -k \frac{e^2}{2r}$$

and is the lower (minus sign!), the smaller the orbital radius.

## 22.14. Many-Electron Atoms.

### Origin of Optical and X-Ray Spectra of Atoms

As in the hydrogen atom in more complex atoms electrons can move around the nucleus only in definite selected orbits. Numerous experimental data indicate that possible orbits of electrons in an atom are grouped in the form of *electron shells*. These shells can roughly be represented as concentric spheres surrounding the nucleus (Fig. 368). Each shell contains a definite number of orbits with a single electron in each orbit. The shell of the minimum radius, known as the *K-shell*, contains two orbits. The second shell, or the *L-shell*, has eight orbits. There is the same number of orbits in the next (third) shell. The fourth shell contains 18 orbits, and so on.

As was mentioned in the previous section, a transition of an electron from a larger to a smaller orbit is accompanied by the liberation of energy. The electron occupying the outer shell will necessarily “jump” to the inner shell if only it has a vacant orbit. For this reason, if a many-electron atom is in the ground (unexcited) state, all electrons are concentrated on inner orbits.

Let us consider, for example, an element with atomic number 11, viz. sodium. The charge of the nucleus of a sodium atom is  $Ze = 11e$ . The sodium atom contains 11 electrons: ten of these electrons occupy all available orbits in the *K*- and *L*-shells, while the last (eleventh) electron is in the third shell (Fig. 369).

The bond between the outer electrons of an atom and the nucleus is weaker than for the inner electrons. Firstly, they are at a much larger distance from the nucleus. Secondly, the force of attraction of outer electrons by the positive nucleus is compensated to a larger extent by the repulsion

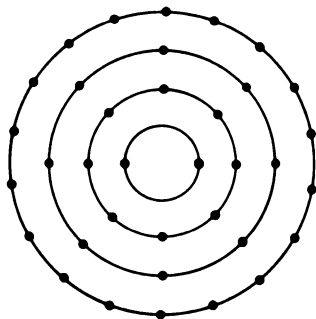


Fig. 368.

Schematic diagram of electron shells in an atom: the number of dark circles is equal to the maximum possible number of electrons in a shell.

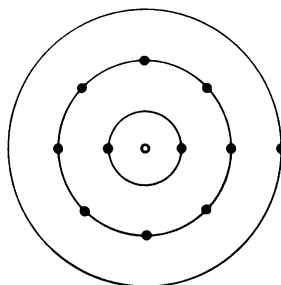


Fig. 369.

Schematic diagram of a sodium atom: light circle is the nucleus, and dark circles, electrons. The *K*- and *L*-shells are filled completely, and one electron is in the third shell.

exerted by the negative electrons occupying the inner shells. Measurements show that depending on the species of atoms, the energy from 5 to 20 eV is required for detaching one of the outer electrons from the atom. In order to transfer one of the outer electrons to a more remote shell without detaching it from the atom (i.e. to excite the atom), a still lower energy is required. When such an electron returns to a shell closer to the nucleus, a light quantum will be emitted whose energy does not exceed 5-20 eV, i.e. with a wavelength in the range of visible or ultraviolet radiation. Therefore, the emission of light in the optical range is associated with the behaviour of the outer electrons of an atom.]

In order to detach inner electrons from an atom, a much higher energy (which rapidly increases with the charge of the atomic nucleus) is required. For example, to tear an electron from the *K*-shell, an energy of about 1.1 keV is required for sodium ( $Z = 11$ ), more than 9 keV for copper ( $Z = 29$ ) and about 70 keV for tungsten ( $Z = 74$ ). A transition of electrons from the *L*-shell and the shells following it to a vacant place in the *K*-shell is therefore accompanied by the emission of quanta with a higher energy (smaller wavelength) corresponding to X-ray.

It was mentioned earlier that X-rays are a form of an electromagnetic radiation caused by an abrupt deceleration of electrons in a substance (bremsstrahlung). It is clear now that there also exists another mechanism of emission of X-rays, which can be described as follows. The bombardment of the anode in an X-ray tube by electrons results in the detachment of electrons from the inner shells of atoms constituting the anode. The vacancies formed are occupied by the electrons from the outer shells of the same atoms. Such transitions are accompanied by the emission of X-rays which is referred to as the *characteristic X-rays* for a given atom.

Thus, the emission of X-rays by atoms is associated with its inner electron shells. The analysis of X-ray spectra has therefore provided a valuable information about the structure of inner electron shells of atoms.

### 22.15. Mendeleev's Periodic System of Elements

The periodic law of variation of chemical properties of elements, discovered by D. I. Mendeleev, reflects deep-rooted regularities in atomic structure. For this reason, it is of primary importance not only for chemistry but also for physics. An adequate theory of atomic structure must be in agreement with Mendeleev's law, i.e. must explain the regularities in chemical properties of element reflected in Mendeleev's Periodic Table. Let us find out how the planetary model of atom can tackle this problem.

Chemical properties are manifested in atomic collisions leading to the formation of molecules. But atoms approach one another to close distances during collisions so that their electron shells interact. It follows hence that

Groups of elements

|   | I | II | III | IV | V | VI | VII | VIII |
|---|---|----|-----|----|---|----|-----|------|
| 1 |   |    |     |    |   |    |     |      |
| 2 |   |    |     |    |   |    |     |      |
| 3 |   |    |     |    |   |    |     |      |

1st shell (2 electrons)  
2nd shell (8 electrons)  
3rd shell (8 electrons)

Fig. 370.

The first three periods of the Mendeleev Periodic Table. The arrangement of electrons in atomic shells is given for all elements.

the chemical properties of an atom are determined by the structure of its electron shells, i.e. ultimately by the charge of the atomic nucleus.<sup>24</sup> For this reason, the elements in the Periodic Table are arranged in the order of increasing nuclear charge. (This also explains identical chemical properties of isotope atoms whose nuclei differ in mass but have equal charges.)

Figure 370 represents the first three periods of the Periodic Table with the arrangement of electrons on possible orbits for each element. As was mentioned in the previous section, possible (allowed) orbits are grouped in shells (*X*, *L*, etc.).

It is noteworthy that the number of a group in Mendeleev's Periodic Table containing a given element is equal to the number of electrons in the *last* of the filled atomic shells.<sup>25</sup> For example, the first group contains hydrogen (with one electron in the *K*-shell), lithium (with one electron in the *L*-shell), sodium (with one electron in the third shell), etc. All these elements possess similar chemical properties. The second group comprises beryllium (two electrons in the *L*-shell), magnesium (two electrons in the third shell), and so on. The second-group elements are also similar in chemical properties. A similar situation takes place for the remaining groups. Hence it follows that the chemical properties of atoms are determined by the electrons in the last shell which is not filled completely. These electrons

<sup>24</sup> It should be recalled that the total number of electrons in the electron shells of an atom is equal to the charge of the atomic nucleus (in units of elementary charge). In turn, the charge is equal to the atomic number of the element in the Periodic Table.

<sup>25</sup> Except for the atoms with completely filled last shell, i.e. the atoms whose electrons occupy all available places in a shell. These atoms constitute the zeroth group (see below).

are known as *valence electrons*. The number of valence electrons determines the valence of the element. For example, alkali metals (Li, Na, K, Rb and Cs) which have *one* valence electron each are *monovalent*. All alkaline-earth elements (Mg, Ca, Sr and Ba) are *bivalent* and have *two* valence electrons, and so on. Atoms with completely filled electron shells do not have any valence electrons and are chemically inert. These are inert gases, viz. helium, neon, argon, etc. which constitute the zeroth group since their valence is zero.

As the number of electrons in atoms increases, the properties of elements change from metallic to nonmetallic. When the next shell is completely filled by electrons, we have an inert gas. With a further increase in the number of electrons, a new electron shell begins to be filled, thus opening the next period in the Periodic Table, in which the transition from metals to nonmetals takes place again.

Starting with the fourth period, a deviation from the above-indicated order of filling electron shells is observed in the Periodic Table. In some regions, a new shell begins to be filled even before the construction of the previous shell has been completed. On other regions, the number of electrons in the last shell remains unchanged when the number of electrons increases, and the previous shells are being completed instead. In such a case, a group of neighbouring elements is formed with the same number of valence electrons, i.e. with similar chemical properties. Rare-earth elements can serve as an example of such a group.

Thus, we have established the reason behind the periodicity of chemical properties of elements. It is due to the fact that the chemical properties of elements are mainly determined by the number of outer (valence) electrons in an atom, while the number of outer electrons is periodically repeated as the *K*-, *L*-, etc. shells are being filled.

But why only outer electrons and not the entire set of the electrons in an atom determine the chemical properties of the atom? As a matter of fact, the energy liberated or absorbed in chemical reactions does not exceed a few electron volts per atom (see Ex. IV. 3). This energy is sufficient for changing the arrangement of the outer electrons of the atom. However, this energy is too low to change the orbits of inner electrons for which the energy of transitions is much higher (see Sec. 22.14). For this reason, when atoms combine into molecules, the arrangement of inner electrons of the atoms being combined is retained. This is proved by the fact that the X-ray emission spectrum for chemical compounds (which is induced, for example, by the electron bombardment) is a superposition of the emission spectra of the elements constituting a compound.

In contrast to the X-ray spectrum, the optical spectrum is determined, as was mentioned earlier, by the behaviour of the outer electrons, i.e. the electrons that determine the chemical properties of the atom as well. For this reason, similar elements are characterized by similar optical spectra.

When a molecule is formed from atoms, the “chemical” (valence) electrons, which are also the “optical” electrons, are regrouped. Consequently, the formation of a molecule is accompanied by a change in the optical properties of atoms. That is why the optical spectrum of a molecule normally differs sharply from the spectra of the atoms constituting the molecule.

Before concluding this section, let us consider the stability (strength) of atoms mentioned at the beginning of this chapter. It is due to the stability of atomic nuclei. The changes in the properties of an atom (like its ionization or formation of compound molecules from atoms) are confined to the regrouping of outer electrons and do not affect the atomic nucleus. Therefore, after such changes the atom can easily be reconstructed (neutralization of ions, decomposition of a molecules, and so on). However, in order to radically change the properties of an atom, as a result of which the atom may change its nature so that its reconstruction involves a new complex process, the charge of the nucleus must be changed. This causing a change in the normal number of electrons in the atom associated with it. In general, it is possible to change the charge of the nucleus. But in view of a very small size and high strength of nuclei, this problem requires special, exceptionally powerful means which will be described in Chap. 24.

### 22.16. Quantum and Wave Properties of Photons

As was mentioned in Sec. 21.3, the laws of photoelectric effect were explained in 1905 by Einstein who used the concept of light quanta (photons). According to his ideas, the energy of an electromagnetic field cannot be divided into arbitrary parts, but rather is emitted and absorbed always in definite portions equal to  $h\nu$ . Here  $\nu$  is the frequency of oscillations in the radiation and  $h$  is Planck's constant. These portions of energy of an electromagnetic field were called light quanta, or photons.

The quantum nature of electromagnetic radiation is usually manifested in experiments when the energy of each photon is sufficiently high and the number of photons is not too large. However, in many optical experiments clearly revealing the wave properties of light, we encounter the opposite situation when the energies of photons are very low but their number is very large (see example set in Sec. 21.3). For this reason, the quantum nature of light remained undiscovered by scientists for a long time.

As was mentioned earlier, it was revealed in the experiments on photoelectric effect for conductors that the maximum kinetic energy of electrons escaping under the effect of light (photoelectrons) is connected with the work function  $A$  and the frequency of electromagnetic waves incident on the conductor through the following relation:

$$h\nu = A + \frac{mv^2}{2}. \quad (22.16.1)$$

In 1916, this relation was confirmed by the American physicist R. Millikan. Fine and thorough measurements made by him in experiments described in Sec. 21.2 have made it possible to establish a linear dependence between the maximum energy received by an electron from light and the frequency of this light, to prove the universal nature of Planck's constant  $h$  and to measure its value ( $h = 6.6 \times 10^{-34}$  J·s). In subsequent experiments, the frequency of radiation incident on the surface of a metal was varied over a wide range from visible light to X-rays. The results of experiments were in excellent agreement with the theory for the entire frequency range under investigation.

In experiments with X-rays, the concepts of quanta were subjected to especially thorough and diversified verification. Indeed, the quanta of visible light (photons) possess a very low energy. For example, for yellow light with  $\nu \approx 5 \times 10^{14}$  s<sup>-1</sup>,  $h\nu \approx 3.31 \times 10^{-19}$  J. Therefore, in most of experiments such a light can be registered only with a large number of photons per unit time. Accordingly, the effect produced by light quanta flying in all directions and having a random distribution can hardly be distinguished from the action of a wave uniformly propagating in all directions. The higher the energy of quanta, the easier the effect of a single quantum can be observed. In other words, it is easier to carry out an experiment for observing the transfer of radiant energy in flashes in definite directions than a uniform energy transfer in all directions. The energy of photons in the X-ray region is much higher than the energy of photons of visible light. Besides, in experiments with X-rays it is easier to realize the conditions for the emission of a small number of quanta per unit time.

In order to obtain X-rays, the anode of an X-ray tube must be bombarded by electrons (see Secs. 17.3 and 17.5). Each act of braking (deceleration) of electrons in the anode material is accompanied by the emission of X-rays. The theory of light quanta predicts that in the most favourable case the entire kinetic energy of an electron stopped in the material is completely converted into a single light quantum whose energy  $h\nu$  is determined from the condition  $W_k = h\nu_{\max}$ . If the electron has been accelerated by a potential difference  $U$ , then  $W_k = eU$ .

Therefore, the maximum frequency of X-rays is specified by the relation

$$h\nu_{\max} = eU. \quad (22.16.2)$$

Indeed, the measurements confirmed that the X-ray spectrum obtained in such experiments is characterized by the wavelength threshold

$$\lambda_{\min} = \frac{c}{\nu_{\max}},$$

where  $c$  is the velocity of light, the maximum frequency of radiation being determined by (22.16.2). Shorter waves (corresponding to larger values of

frequency  $\nu$ ) are never observed here, while longer waves correspond to the conversion of a part of the kinetic energy of an electron into the energy of X-rays. The frequency threshold of the X-ray spectrum can be determined quite liably. For this reason, such experiments were used for determining Planck's constant (with the help of (22.16.2)). The best measurement made by this method yielded  $h = 6.624 \times 10^{-34} \text{ J}\cdot\text{s}^{26}$ . These results are in agreement with those obtained by measuring  $h$  in experiments on photoelectric effect. Thus, the quantum theory is confirmed not only by the results of experiments on absorption of radiant energy (photoelectric effect) but also of experiments on energy emission.

By controlling the number of electrons bombarding the anode of an X-ray tube, we can vary the number of emitted X-ray quanta. If we then expose a metal plate to the emitted X-rays, thus causing the emission of photoelectrons, the experiments show that the kinetic energy of these electrons will be equal to the energy of X-ray quanta (since the energies of electrons and X-ray quanta in these experiments amount to tens of kiloelectron volts, the work function of electrons, which is within a few electron volts, can be neglected).

Thus, the entire cycle of energy conversion in such experiments is as follows: (1) the conversion of the work  $eU$  done by the electric field into the kinetic energy  $W = mv_e^2/2$  of an electron in the X-ray tube, (2) the conversion of the kinetic energy of an electron into the energy of an X-ray quantum emitted by it as a result of an abrupt deceleration and (3) the absorption of a photon by an electron and the conversion of the photon energy into the kinetic energy of the photoelectron:

$$eU = \frac{1}{2} mv_e^2 = h\nu = \frac{1}{2} mv^2.$$

Such experiments can be varied by taking advantage of convenient conditions provided by experiments with X-rays. All of them show that the energy in these phenomena is transferred in concentrated portions rather than accumulated gradually as would be the case if energy were transferred continuously in the form of an electromagnetic wave. One of the most convincing experiments of this kind was made by A. F. Ioffe (1880-1960). Direct experiments on detection of individual photons were also carried out. They showed that the energy of X-rays propagates from the anode of the tube not simultaneously in all directions but in the form of portions (quanta) emitted one by one in different directions.

Thus, the analysis of photoelectric effect and experiments with X-rays convincingly proved that light behaves in these phenomena not as a wave

---

<sup>26</sup> The most accurate value of Planck's constant obtained in latest experiments is  $h = (6.626\,176 \pm 0.000\,036) \times 10^{-34} \text{ J}\cdot\text{s}$ .



but as a particle, viz. photon, which is formed during emission, flies in a certain direction and gives away its entire energy to another particle as a result of absorption. But if a photon behaves as a particle with a total energy  $W = h\nu$ , it must have a certain momentum. The photon has a velocity equal to the speed of light. Therefore, from the general formulas of relativistic mechanics (see Secs. 22.6 and 22.7) it can be expected that it will have a momentum

$$p = \frac{\nu}{c^2} W = \frac{1}{c} W = \frac{h\nu}{c}. \quad (22.16.3)$$

As was shown earlier (see Sec. 22.7), the distinctive feature of a photon is that its rest mass is zero: the photon always moves at the velocity of light and cannot exist at rest.

The fact that photons possess momentum implicitly follows even from experiments on light pressure (see Sec. 7.1). The ability of light to exert a pressure on a reflecting or an absorbing surface should be interpreted as a result of transfer of the momentum of photons by analogy with the gas molecules reflected by the vessel wall, which impart a momentum to the wall and exert a pressure on it (see Vol. 1).

An important role in the development of the concept of photons as elementary particles was played by the experiments carried out by the American physicist Arthur Compton (1892-1962), in which it was shown directly that photons colliding with electrons behave as particles having an energy and a momentum connected through relation (22.16.3).

By analyzing the scattering of X-radiation in a substance composed of light atoms (Fig. 371), Compton observed (in 1923) that in this process the wavelength of X-rays changes and established the following relation between the change  $\Delta\lambda$  in the wavelength and the scattering angle  $\theta$ :

$$\Delta\lambda = 2\lambda_0 \sin^2 \frac{\theta}{2}. \quad (22.16.4)$$

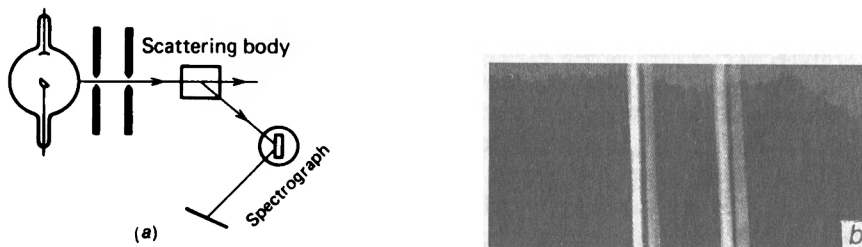


Fig. 371.

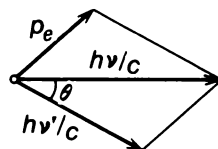
(a) Schematic diagram of Compton's experiment. (b) The spectrum of scattered X-radiation.

The constant  $\lambda_0 = h/m_e c = 2.43 \times 10^{-12} \text{ m}$  appearing here was initially determined from experiments. The results of these experiments contradict to the classical ideas about the scattering of electromagnetic waves by atoms, according to which an atom should perform forced oscillations under the action of incident radiation and should become a source of scattered waves of the same frequency (i.e. the same wavelength) as the incident wave.

The phenomenon discovered by Compton was, however, brilliantly interpreted with the help of the concept of quanta. Compton's experiments were carried out with X-ray quanta having an energy of 17.5 keV. This energy is high in comparison with the binding energy of electrons in light atoms (which is of the order of several electron volts). Therefore, we can assume that in Compton's experiments a photon collides with a free electron (and not with an atom as a whole), the collision resembling that of elastic balls.

Fig. 372.

Elastic collision between a photon and an electron. The electron is at rest before the collision:  $h\nu/c$  is the momentum of the incident photon,  $h\nu'/c$  is the momentum of the scattered photon,  $p_e$  is the electron momentum and  $\theta$  is the scattering angle.



Using the energy and momentum conservation laws (Fig. 372) for this collision, we obtain

$$h\nu + m_e c^2 = h\nu' + \sqrt{m_e^2 c^4 + p_e^2 c^2},$$

$$p_e^2 = \left(\frac{h\nu}{c}\right)^2 + \left(\frac{h\nu'}{c}\right)^2 - 2 \frac{h^2}{c^2} \nu\nu' \cos \theta. \quad (22.16.5)$$

While determining  $p_e^2$ , we must take into account the vector nature of the momentum conservation law and make use of the trigonometric theorem about the relation between the lengths of the sides of a triangle (see Fig. 372):

The *recoil electrons* that have received during the scattering of high-energy X-ray photons a momentum from these photons may have velocities comparable with the speed of light. For this reason, we must take into consideration the relativistic increase in their mass and use the laws of relativistic mechanics (see Secs. 22.6 and 22.7) as was done in (22.16.5). After some transformation, the solution of the system of equations (22.16.5) leads to a quantitative explanation of relation (22.16.4) for the Compton effect that was earlier established experimentally (see Ex. IV.19). In subsequent experiments with high-energy quanta, the Compton scattering was observed not only during interactions with electrons but also with other particles (say, protons and neutrons). Thus, these experiments directly revealed that a photon behaves as an elementary particle not only in phenomena involving

photoelectric effect and emission but also in interactions with electrons and other particles.<sup>27</sup>

Further experiments confirmed the idea that a photon is a particle. Processes have been discovered in which a photon perishes during interactions with atomic nuclei, giving birth to a pair of elementary particles, viz. electron and positron (a particle having the same mass as an electron and a positive charge equal in magnitude to the electron charge), the nucleus remaining unchanged in such process (see Sec. 24.6). These experiments proved that electrons and positrons are not emitted by the nucleus since the latter does not change, but emerge under the action of light. An electron, a positron and a nucleus flying apart have energies and momenta borrowed from the vanishing photon.

A reverse process has also been discovered in which interacting electron and positron cease to exist as elementary charged particles. The charges of the electron and positron are neutralized, and their rest energy is converted into the energy of a photon pair formed in such a process and flying apart at the velocity of light.

It will be shown later (see Chap. 25) that such mutual conversions of some elementary particles into others constitute their important and peculiar property. In this respect, a photon does not differ from other microparticles like electrons or protons.

Finally, it should be noted that photons, like all other particles, may experience the action of a gravitational field. For example, accurate measurements made during total solar eclipses and concerning the positions of stars, the light from which passes near the Sun, indicate that this light experiences the attraction of the Sun and is deflected from its initial path. Qualitatively, this can be explained by taking into account the fact that photons possess an energy  $h\nu$  which corresponds to the "mass due to motion"  $m = h\nu/c^2$  that experiences the gravitational attraction to the Sun. Another experimentally observed beautiful effect consists in that a photon moving in a gravitational field changes its energy. For example, the photon energy  $W = mc^2 = h\nu$  changes as a result of its motion in the gravitational field of the Earth due to the change in the potential energy of the photon in this field by

$$mgH = \frac{h\nu}{c^2} gH,$$

where  $H$  is the distance covered by the photon in the direction of the gravitational field of the Earth. Hence we may conclude that the photon frequency

---

<sup>27</sup> From the point of view of the modern theory of elementary particles, the Compton scattering is interpreted as the absorption of a photon  $h\nu$  by an electron (or other particle), which is followed by the emission of a new particle, viz. a photon  $h\nu'$ .

changes by

$$\Delta\nu = \frac{\nu gH}{c^2}.$$

In experiments involving the motion of photons emitted by excited atomic nuclei in the gravitational field starting with height  $H = 22.6$  m to the ground, the change in the photon frequency was registered, which coincided with the theoretical predictions:

$$\frac{\Delta\nu}{\nu} = \frac{gH}{c^2} = \frac{9.8 \text{ m/s}^2 \times 22.6 \text{ m}}{9 \times 10^{16} (\text{m/s})^2} = 2.5 \times 10^{-15}.$$

Thus, the fact that photons are subjected to gravitational effects has been confirmed.

Having considered numerous experiments, we arrive at the conclusion that in some cases light must be regarded as a flow of corpuscles, viz. photons, having the properties typical of other microparticles. However, such phenomena as interference and diffraction can only be explained by the wave properties of electromagnetic radiation. These two aspects of nature, i.e. wave and corpuscular, turn out to be equally important. For this reason, in order to explain all peculiarities in the behaviour of radiation, it is necessary to admit that under certain conditions electromagnetic waves have the properties of flows of particles. The inverse statement that the particles of electromagnetic field (photons) exhibit wave properties is equally justified. Such a wave-particle duality of photons contradicts to the formed classical ideas about waves and particles as completely different concepts.

It seemed initially that photons possessing such peculiar properties essentially differ from other particles like electrons or protons. The further development of physics of microcosm, however, has made it possible to establish that the wave-particle duality is not at all a specific feature of photons but is of a more general nature.

## 22.17. Fundamentals of Quantum (Wave) Mechanics

The investigation into atomic structure leads to the conclusion that the behaviour of electrons in an atom, like the behaviour of photons, cannot be described by the laws of classical physics, i.e. the laws established in experiments with macroscopic bodies. The existence of discrete energy levels in electron shells of an atom and the regularities governing the transitions between the levels and the filling of these energy states also cannot be interpreted by using conventional concepts of mechanics and the laws of electromagnetism.

An important step in clarifying these contradictions was made in 1923 by the French physicist Louis de Broglie (b. 1892) who proposed and sub-

stantiated the idea that not only photons but also all particles have wave properties. These properties cannot be explained by the classical laws, but play an important role in atomic phenomena.

Quantum of electromagnetic radiation, viz. photons, are known to be characterized by the momentum  $p = h\nu/c$ . At the same time, a light wave of frequency  $\nu$  has a wavelength  $\lambda = c/\nu$ . Cancelling out the frequency from these expressions, we obtain the relation between the wavelength and the momentum of a photon:

$$\lambda = \frac{h}{p}. \quad (22.17.1)$$

If the properties of photons and other particles are indeed identical as predicted by the hypothesis about the wave-particle duality, this relation must be applicable to any particle. In this way, the formula for the de Broglie wavelength was obtained, i.e. for the wavelength which must be ascribed to a particle having a momentum  $p$  in order to explain its wave properties. This formula has the same form as (22.17.1). If the velocity of a particle with a rest mass  $m$  is small in comparison with the speed of light, the formula for the de Broglie wavelength can be written as

$$\lambda = \frac{h}{mv}. \quad (22.17.2)$$

The validity of the hypothesis put forth by de Broglie was verified in experiments on electron scattering by crystals.

Previously, the scattering of X-rays by crystals has been used to prove the wave nature of X-rays (see Sec. 17.6). Scattering occurs only at certain angles to the incident beam rather than in all directions due to the interference of the secondary waves emitted by regularly arranged atoms of a crystal. In addition to the central spot from a straight beam, a system of rings

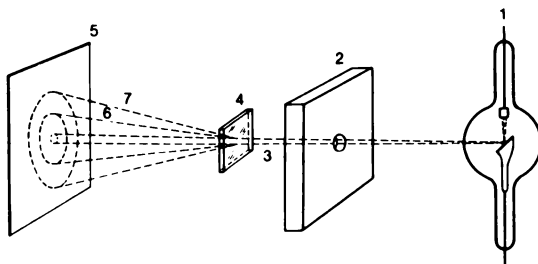


Fig. 373.

Schematic diagram of the experiment on the diffraction of X-rays by crystals: 1 — X-ray tube, 2 — lead diaphragm isolating a narrow beam 3 of X-rays, 4 — polycrystalline sample, 5 — photographic film (wrapped in black paper), 6 and 7 — the beams of X-rays scattered by the crystal.

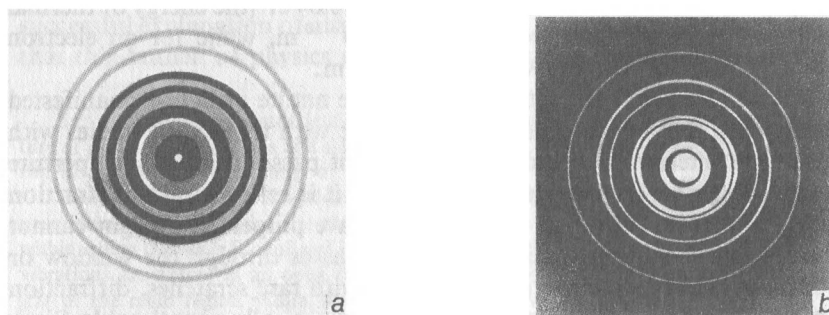


Fig. 374.

Photographs of the diffraction of X-rays (a) and electrons (b) by polycrystalline gold.

from scattered (diffracted) radiation is formed on a photographic film arranged behind the scattering crystal (Fig. 373). An example of such a photograph is shown in Fig. 374a.<sup>28</sup>

It turned out that if the crystal is bombarded by electrons instead of X-rays, scattered electrons also form on the film a system of rings similar to those produced by scattered X-rays (Fig. 374b). Thus, we can draw a remarkable conclusion: *electrons interfere, i.e. possess wave properties.*

Later, diffraction phenomena were observed for other particles, i.e. atoms, molecules and neutrons.<sup>29</sup> These experiments irrefutably proved that in some phenomena the microparticles behave like waves. Experiments also allowed scientists to determine the wavelength that must be ascribed to a given particle in order to explain its diffraction. The results were in complete agreement with the de Broglie formula (22.17.2): *the wavelength turned out to be inversely proportional to the product of the mass  $m$  of a particle and its velocity  $v$ , the proportionality factor being Planck's constant.*

Planck's constant is very small:  $h = 6.6 \times 10^{-34} \text{ J}\cdot\text{s}$ . For this reason, the de Broglie wavelength for a particle of an appreciable mass is negligibly small. According to the de Broglie formula, a dust particle with a mass of a microgram ( $10^{-9} \text{ kg}$ ) and flying at a velocity of  $1 \text{ cm/s}$  has a wavelength  $\lambda = 6.6 \times 10^{-34} / (10^{-9} \times 10^{-2}) = 6.6 \times 10^{-23} \text{ m}$ . This value is negligibly small even in comparison with atomic dimensions.

A different situation takes place for electrons or atoms with a mass much smaller than a microgram. At a moderate velocity, the wavelength corresponding to them is of the order of the wavelength of X-rays. For

<sup>28</sup> Figure 374 shows the patterns obtained with a polycrystalline sample, i.e. a crystal consisting of a large number of small randomly oriented crystals. With such a sample, individual spots of scattered radiation merge into rings surrounding the central spot.

<sup>29</sup> A neutron is a neutral microparticle with a mass of about 1 amu (see Secs. 24.2 and 24.3).

example, for a helium atom with an energy of 0.04 eV (the energy of thermal motion at room temperature),  $\lambda = 0.7 \times 10^{-10}$  m, while for an electron with an energy of 13.5 eV,  $\lambda = 3.3 \times 10^{-10}$  m.

It is well known from optics that the wave nature of light is manifested clearly when the wavelengths are comparable with the size of bodies with which light interacts. For example, when light passes through an aperture having a size of a few wavelengths, or when it is reflected at a diffraction grating with a small grating constant, the wave properties of light cannot be neglected. On the contrary, when light passes through the window or a room, or when it is reflected at a mirror with rare scratches, diffraction phenomena can be disregarded since they are practically unnoticeable. Similarly, *the wave properties of particles are important only when the de Broglie wavelength is not small in comparison with the size of objects with which interaction takes place.* During interactions of atoms with electrons and other microparticles for which the de Broglie wavelength is of the order of atomic dimensions, the wave properties of particles play a significant and sometimes decisive part. The more so it refers to the processes associated with the behaviour of electrons within atoms or molecules.

When particles of macroscopic dimensions interact, for which it was shown that the de Broglie wavelength is about  $10^{-9}$  of their size, the wave properties should not be taken into account. For this reason, classical mechanics, whose laws are derived from observations of large bodies and never take into account the wave properties of bodies, gives good results in problems involving the motion of celestial bodies, parts of machines, etc. But just for this reason classical mechanics is completely inapplicable for interpreting atomic phenomena. Problems in atomic physics cannot be solved with the help of Newtonian mechanics. A new, more perfect, mechanics had to be worked out, which would take into account the wave properties of matter.

This important problem has been solved by the end of the twenties. The main contribution to its solution was made by the German physicist Werner Karl Heisenberg (1901-1976), the Austrian physicist Erwin Schrödinger (1887-1961) and the English physicist Paul Adrien Maurice Dirac (1902-1984).

The system of laws of motion for particles of a substance, which takes into account their wave properties, is known as *wave*, or *quantum mechanics*. Quantum mechanics has solved a large number of problems associated with the behaviour of particles of the microcosm. They include the behaviour of electrons in atoms and molecules and the interaction between atoms, followed by the emission and absorption of light, the collisions of electrons and other particles with atoms, ferromagnetism and many other phenomena. Quantum mechanics has also predicted a number of new

phenomena. All its predictions have been confirmed by experiments. The successful explanation of atomic phenomena by quantum mechanics proves that this branch of physics correctly reflects the objective laws of nature.

Let us consider in greater detail some questions concerning the quantum-mechanical nature of phenomena in atoms and show how the formulas describing the energy levels of atoms can be obtained with the help of wave concepts.

The electric field of the nucleus keeps an electron in an atom in a certain region of space near the nucleus. Regarding the electron as a wave, we cannot indicate precisely the volume within which this wave is confined just as we cannot indicate a sharp boundary beyond which vibrations are absent in an open tube (see Fig. 107). By the size of an atom we mean the size of the main region within which the electron wave is confined.

Consistent wave concepts concerning the behaviour of an electron in an atom can be formulated by using the laws of quantum mechanics. Quantum-mechanical calculations make it possible, in particular, to determine definite states of an atom and to specify the discrete energy levels corresponding to these states. However, quantum-mechanical laws are expressed by rather complicated mathematical formulas and are beyond the scope of this book.

Nevertheless, some corollaries from these laws can be established on the basis of the concept of the de Broglie wave. By way of an example, let us consider the hydrogen atom.

It should be recalled that while considering the planetary model of the atom (see Sec. 22.13), we mentioned that an electron can move around the nucleus in some allowed orbits. Although in quantum mechanics, where an electron is treated as a certain wave, we cannot speak about the motion in an orbit, we shall still use the concept of electron "orbit" and take advantage of the de Broglie waves associated with the electron in order to indicate which "orbits" are allowed. Although such an approach is neither consistent nor strict, it is very visual and allows us to obtain the results close to those obtained from exact quantum-mechanical calculations.

Thus, we shall consider the motion of the electron in a hydrogen atom in a circular orbit of radius  $r$ . We shall state that only those orbits are allowed into which an integral number of the de Broglie wavelengths can fit, i.e. the orbits for which

$$2\pi r = \lambda n \quad (n = 1, 2, 3, \dots). \quad (22.17.3)$$

If this condition (which is known as the *condition of orbit quantization*) is satisfied, any point on an orbit corresponds to a certain phase of oscillations associated with a wave.

Indeed, having specified a point on an orbit, we note that after completing a turn in the orbit the wave arrives at this point in the same phase.

Thus, when the condition of orbit quantization is satisfied, the wave pattern becomes definite and unambiguous. On the contrary, if the condition is violated, the wave will arrive at the specified point after completing a turn in a different phase, then again in a new phase, and so on. In other words, we have no unambiguous wave pattern in this case. Therefore, the wave motion of electrons in a bounded space is reduced, like for other wave phenomena, to the formation of "standing waves" (see Sec. 5.5-5.8, 6.3 and 6.6).

These standing waves satisfy the "boundary condition" (22.17.2) which connects the kinetic energy  $mv^2/2$  of the electron with the atomic size. Indeed, using the de Broglie formula  $\lambda = h/mv$ , we obtain

$$\frac{mv^2}{2} = \frac{h^2}{2m\lambda^2} = \frac{h^2 n^2}{8m\pi^2 r^2}. \quad (22.17.4)$$

The potential energy of the electron on an orbit is  $W_p = -ke^2/r$  see (22.13.1). The total



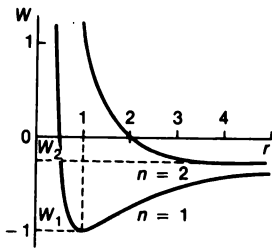


Fig. 375.

The energy of an atom as a function of its size. The curves represent the function (22.17.5) for  $n = 1$  and  $n = 2$ . The energy in the  $k^2 \cdot 2\pi^2 me^4 / h^2$  units is plotted along the ordinate axis, and the atomic radius in the  $h^2 / (k \cdot 4\pi^2 me^2)$  units, along the abscissa axis. In these units, formula (22.17.5) becomes  $W = n^2/r^2 - 2/r$ . Dashed lines indicate the energy minima corresponding to the ground  $W_1$  and the first excited  $W_2$  state of the hydrogen atom.

energy of the atom, i.e. the sum of the kinetic and potential energies of the electron, is

$$W = \frac{mv^2}{2} + W_p = \frac{h^2 n^2}{8m\pi^2 r^2} = -k \frac{e^2}{r}.$$

It is convenient to represent this expression in the form

$$W = \left( \frac{nh}{2\sqrt{2}m\pi r} - k \frac{e^2 \pi \sqrt{2}m}{nh} \right)^2 - k^2 \frac{2\pi^2 me^4}{h^2 n^2}. \quad (22.17.5)$$

The curve representing the dependence of  $W$  on  $r$  is shown in Fig. 375. As the size of the atom decreases, its energy also decreases, passes through the minimum and then increases. The atom will be in a stable state when its size corresponds to the minimum energy. Indeed, in this case, any change in the size of the atom requires an expenditure of energy and cannot occur spontaneously.

Energy  $W$  passes through the minimum for the values of  $r$  that nullify the positive definite term, viz. the square of the expression in the parentheses in (22.17.5). Thus, for stable states of the atom the energy is given by

$$W_n = -k^2 \frac{2\pi^2 me^4}{h^2 n^2} = - \left( \frac{1}{4\pi\epsilon_0} \right) \frac{2\pi^2 me^4}{h^2 n^2}, \quad (22.17.6)$$

where  $\epsilon_0$  is the electric constant equal to  $8.85 \times 10^{-12}$  F/m,  $m$  is the electron mass and  $n = 1, 2, 3, \dots$  is the *principal quantum number* indicating the number of an energy level. The value  $n = 1$  corresponds to the minimum energy of the atom.

It should be noted that a rigorous quantum-mechanical solution of the problem on the energy levels of the hydrogen atom leads to a result coinciding with (22.17.6).

The system of energy levels of the hydrogen atom defined by (22.17.6) coincides with the one shown in Fig. 360 (for the origin of energy in Fig. 360, we take the ground state of the atom with  $n = 1$ , i.e. the constant  $k^2(2\pi^2 me^4)/h^2$  is added to (22.17.6)).

The exact solution also indicates that only the ground state of the atom corresponding to the lowerest energy level ( $n = 1$ ) is stable. The remaining states ( $n > 1$ ) turn out to be not quite stable. After some time, they go over to lower energy states by emitting the excess energy in the form of a light quantum.

It becomes clear now why atoms are stable, i.e. why an electron cannot fall on the nucleus. This is prevented by a rapid increase in the kinetic energy of the electron, which accompanies the decrease in the de Broglie wavelength corresponding to it, as a result of reduction in the atomic size see (22.17.4).

It should be noted once again that quantum mechanics does not contradict Newtonian classical mechanics. All conclusions of Newtonian mechanics are contained in quantum mechanics and can be derived from

it as approximate solutions which are completely applicable when the wave properties of particles do not play a significant role. A similar situation is found in the theory of relativity (see Secs. 22.6 and 22.7) which coincides with Newtonian mechanics when the velocities of particles are small in comparison with the speed of light. In atomic physics, one often comes across phenomena in which wave properties are significant and the velocities of particles are high. In such cases, one should take into account both the quantum theory and the theory of relativity, i.e. to make use of relativistic quantum mechanics.

It should also be mentioned that modern physics has already dealt with problems that cannot be solved completely even with the help of relativistic quantum mechanics. This concerns some properties of atomic nuclei, as well as interactions and properties of nuclear particles. A further improvement of quantum mechanics, that has not been completed so far, is required for this purpose.

? **IV.1.** Having passed across a potential difference of 1000 V, a particle acquired an energy of 8 keV. What is the particle charge?

**IV.2.** Determine the velocity of a helium atom whose kinetic energy is 2 MeV.

**IV.3.** Determine the energy (in electron volts) liberated during the formation of a  $\text{CO}_2$  molecule from carbon and oxygen if the heat of formation of  $\text{CO}_2$  is 395 kJ/mole.

**IV.4.** An ion describes a circle of radius 5 cm in a magnetic field. What will be the radius of the trajectory of an ion with four times the mass, the same charge and (a) the same velocity of (b) the same energy?

**IV.5.** The radius of the trajectory of an  $\text{He}^+$  ion ( $q/m = e/4m_u$ , where  $e$  is the elementary charge and  $m_u = 1$  amu) is 10 cm. Determine the radius of the trajectory of a particle with twice the charge-to-mass ratio in the same magnetic field if it has been accelerated by the same potential difference. Analyze the following cases:

$$(a) \quad \frac{q}{m} = \frac{2e}{4m_u} \quad (\text{He}^{++} \text{ ion})$$

and

$$(b) \quad \frac{q}{m} = \frac{e}{2m_u} \quad (\text{H}_2^+ \text{ ion}).$$

**IV.6.** Calculate the radius of the trajectory of a singly charged ion with a mass of 20 amu in a magnetic field of 0.05 T if the ion has been accelerated by a potential difference of 1000 V.

**IV.7.** Two identical particles having different velocities move in circular trajectories in the same magnetic field. Which particle completes one revolution faster?

**IV.8.** Write the expression for the time required for a charged particle to complete one revolution in a magnetic field. Calculate the time of one revolution for a particle of a charge  $e$  and a mass of 1 amu in a magnetic field of 1.5 T.

**IV.9.** What are the ratios of the masses of a moving electron and a hydrogen atom to the corresponding rest masses if the kinetic energy of the particles is (a) 1 keV, (b) 1 MeV and (c) 1 GeV?

**IV.10.** A projectile of 1-kg mass moves at a velocity of 1000 m/s. Determine the additional mass of the projectile due to the motion.

**IV.11.** Prove that the classical expression  $W = m_0 v^2/2$  follows from the relativistic formula (22.6.2) for kinetic energy in the case of velocities ( $v/c \ll 1$ ).

*Hint.*

$$1 - \sqrt{1 - v^2/c^2} = \frac{v^2 c^2}{1 + \sqrt{1 - v^2/c^2}} \approx \frac{1}{2} \frac{v^2}{c^2}.$$

**IV.12.** Using formula (22.6.2) for the dependence of mass on velocity, determine the ratio of the mass of a moving body to the rest mass for velocities such that  $v/c = 0.1, 0.99$ .

**IV.13.** What will be the increment in the mass of 1 g of helium due to heating from  $0^\circ$  to  $1000^\circ\text{C}$  (the heat capacity of 1 g of gaseous helium at constant volume is  $3.12\text{ J/K}$ )?

**IV.14.** Determine the distance between the centres of the slits of receivers in a mass spectrograph (see Fig. 354) intended for separating uranium isotopes  $^{238}\text{U}$  and  $^{235}\text{U}$  if the radius of the trajectory of  $^{235}\text{U}$  ions is 50 cm.

**IV.15.** The current of the  $\text{U}^+$  ion beam in a mass spectrograph is 1 mA. What amount of  $^{235}\text{U}$  will precipitate in the receiver over a day?

**IV.16.** The glow of atomic hydrogen is excited by electrons with an energy of 12.5 eV. What is the energy of emitted quanta? Use the energy level diagram for the hydrogen atom (see Fig. 360).

**IV.17.** Using the energy level diagram in Fig. 360, determine the wavelengths in the absorption spectrum of atomic hydrogen.

**IV.18.** Calculate the electrostatic (Coulomb) and gravitational forces of interaction of the electron with the nucleus in a hydrogen atom. The radius of the atom should be taken as  $0.5 \times 10^{-8}\text{ cm}$ . The gravitational constant is  $6.7 \times 10^{-11}\text{ N}\cdot\text{m}^2/\text{kg}^2$ .

**IV.19.** Using the energy and momentum conservation laws for the elastic Compton scattering of photons by electrons (22.16.5), derive expression (22.16.4) for the change in the wavelength of a photon.

**IV.20.** Calculate the wavelength of a light quantum with an energy of 10 eV and the de Broglie wavelength of an electron and a proton with the same energy.

**IV.21.** Explain why diffraction phenomena are clearly observed during the scattering of slow neutrons (with an energy of the order of 0.01 eV) by crystals (scattering takes place only in certain directions), while for fast neutrons with an energy of the order of 100 eV these phenomena are imperceptible.

## Chapter 23

# Radioactivity

### 23.1. Discovery of Radioactivity.

#### Radioactive Elements

Thorium, uranium and some other elements continuously emit invisible radiation without any external effects (i.e. due to internal factors). This radiation, like X-rays, penetrates through opaque screens and produces photographic and ionization effects.

The property of spontaneous emission of such a radiation is known as *radioactivity*. The elements possessing this property are referred to as *radioactive elements*, and the radiation emitted by them is called *radioactive radiation*. Radioactivity of uranium was discovered in 1896 by the French physicist Antoine Henri Becquerel (1852-1908).

Radioactivity was discovered soon after the discovery of X-rays. The emission of X-rays was first detected during the bombardment of the glass walls of a gas-discharge tube by cathode rays. The most effective result of this bombardment is the intense green glow of glass, viz. luminescence (see Vol. 2, Sec. 8.12). This suggested the idea that X-rays are the product of luminescence and that they accompany any luminescence, for example, the one excited by light.

Becquerel experimentally verified this hypothesis. He exposed luminescent materials to light and then brought them to a photographic plate wrapped in a black paper. The emission of penetrating radiation had to be detected from the blackening of the photographic plate after its development. From all luminescent materials tested by Becquerel, only uranium salts caused the blackening of the plate. However, it turned out that a sample preliminarily excited by intense light produced the same blackening as an unexcited sample. It follows hence that the emission of radiation by a uranium salt is not associated with luminescence and does not depend on external effects. This conclusion was confirmed in experiments with non-luminescent uranium-containing compounds all of which emit penetrating radiation.

After Becquerel's discovery of radioactivity of uranium, the Polish-born French physicist Marie Curie (1867-1934), who made most of her researches together with her husband Pierre Curie (1859-1906), investigated most of the known elements and many compounds in order to establish whether they had radioactive properties. In her experiments, M. Curie used as an indicator of radioactivity the property of radioactive materials to ionize air. This method is much more sensitive than that based on the darkening of a photographic plate. The ionization effect of a radioactive sample can

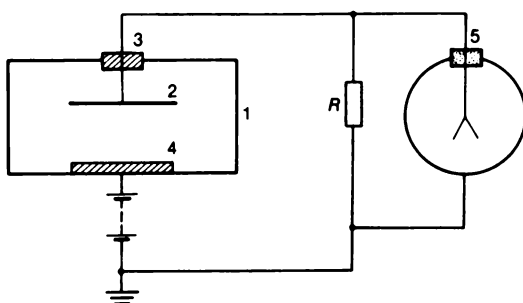


Fig. 376.

Measurement of the ionization current: 1 — ionization chamber casing, 2 — electrode separated from 1 by an insulating plug 3, 4 — sample under investigation, 5 — electrometer. Resistance  $R = 10^8\text{--}10^{12} \Omega$ . When the voltage of the battery is high enough, all the ions formed in the volume of the ionization chamber by ionizing radiation are accumulated at the electrodes, and the current through the chamber is proportional to the ionization effect of the sample. In the absence of ionizing agents, the air in the chamber does not conduct current, and the current is zero.

easily be detected with the help of the experiment represented in Fig. 376 (cf. Vol. 2, Sec. 8.2). M. Curie's experiments led to the following results.

1. Radioactivity is observed not only for uranium but also for all its chemical compounds. Besides, radioactive properties are revealed for thorium and all its chemical compounds.

2. The radioactivity of a sample of any chemical composition is the same as that of pure uranium or thorium in the same amount as their content in the given sample.

The latter result indicates that the properties of a molecule containing a radioactive element do not affect the radioactivity. Thus, *radioactivity is an inherent property of the atoms of a radioactive element rather than a molecular phenomenon*.

In addition to pure elements and their compounds, M. Curie investigated also some mineral rocks. It turned out that the radioactivity of minerals was determined by the presence of uranium and thorium in them. Some minerals, however, exhibited an unexpectedly high radioactivity. For example, pitchblende exhibited four times as high ionization as that of uranium contained in it.

The high radioactivity of pitchblende could be explained only by an admixture of an unknown radioactive element contained in such a small amount that it could not be detected by chemical analysis. Despite its low content, the element emitted a stronger radiant flux than uranium contained in a larger amount. Consequently, the radioactivity of this element must be many times stronger than that of uranium.

Proceeding from these considerations, Pierre and Marie Curie endeavoured the chemical separation of the hypothetical element from pitch-

blende. The radioactivity per unit mass of the end product was an indicator of the success of chemical operations. This quantity had to increase with the content of the new element in the end product. After many years of hard work, they finally succeeded in obtaining a few tenths of a pure element whose radioactivity exceeded that of uranium by more than million times. The element was called *radium* (i.e. radiant).

According to its chemical properties, radium (Ra) is an alkaline earth metal. Its atomic mass was found to be 226. Proceeding from the chemical properties and the mass of radium, the latter was placed into the empty cell No. 88 of the Periodic Table.

Radium is always present in uranium ores although in very small amounts (about 1 g of radium in 3 t of uranium). For this reason, the extraction of radium is a laborious process. Radium is one of the most expensive and rare metals. It is valuable as a concentrated source of radioactive radiation.

Subsequent investigations carried out by the Curies and other scientists considerably extended the number of known radioactive elements.

It turned out that all elements with an atomic number exceeding 83 are radioactive. They were obtained as small admixtures in uranium, radium and thorium.<sup>1</sup>

Radioactive isotopes of thallium ( $Z = 81$ ), lead ( $Z = 82$ ) and bismuth ( $Z = 83$ ) were found in a similar way. It should be noted that only rare isotopes of these elements admixed to uranium, radium and thorium are radioactive. Ordinary thallium, lead and bismuth are nonradioactive.

Besides elements at the end of the Periodic Table, samarium, potassium and rubidium were also found to be radioactive. The radioactivity of these elements is weak and can be detected with difficulty.

## 23.2. Alpha-, Beta- and Gamma-Radiation.

### Wilson Cloud Chamber.

It was mentioned above that radioactive radiation produces the ionization and photographic effects. These effects are exhibited both by fast charged particles and X-rays which are electromagnetic waves. In order to find out whether a radioactive radiation carries a charge, it is sufficient to apply an electric or a magnetic field to it.

Let us consider the following experiment. A radioactive sample 1 (say, a grain of radium) is placed in an evacuated box (Fig. 377a) in front of a narrow slit in a lead partition 2. A photographic plate 3 is placed on the other side of the slit. After the development, we shall obtain on the plate a dark line which is a shadow image of the slit. Consequently, the lead partition does not transmit radioactive radiation that passes through

<sup>1</sup> Except for elements with atomic numbers 85 and 87, which do not exist in natural form.

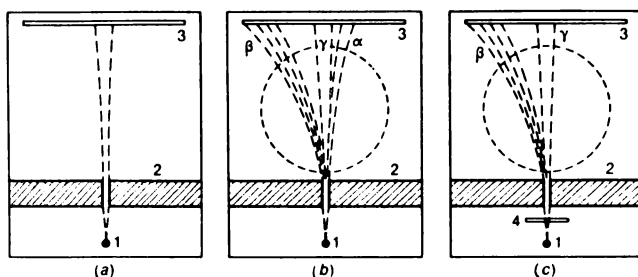


Fig. 377.

Deflection of radioactive radiation by a magnetic field: (a) the trajectories of the rays in the absence of a magnetic field, (b) the trajectories of the rays in a magnetic field (dashed circle is the projection of the magnet poles, the magnetic field lines are perpendicular to the plane of the figure and directed upwards), (c) a 0.1-mm thick sheet of paper completely absorbs  $\alpha$ -radiation. 1 — radioactive sample, 2 — lead screen, 3 — photographic plate, 4 — sheet of paper 0.1 mm thick.

the slit in the form of a narrow beam. Let us now place the box between the poles of a powerful magnet (Fig. 377b) and again place the photographic plate in position 3. Having developed the plate, we shall find not one but three lines on it, the middle line corresponding to the linear propagation of the beam from the sample through the slit.

Thus, the beam of radioactive radiation is split in the magnetic field into three components two of which are deflected in opposite directions, while the third component does not undergo any deflection. The first two components are the flows of oppositely charged particles. The positively charged particles are called  $\alpha$ -particles or  $\alpha$ -radiation. The negatively charged particles are known as  $\beta$ -particles, or  $\beta$ -radiation. Alpha-particles are deflected by the magnetic field to a much smaller extent than  $\beta$ -particles. The neutral component that does not experience any deflection in the magnetic field is called  $\gamma$ -radiation.

Alpha-, beta- and gamma-radiation have different properties, for one, different penetrabilities through a substance. The penetrability of a radioactive radiation can be investigated with the help of the same device (Fig. 377c). We shall place screens of increasing thickness between sample 1 and the slit, expose the photographic plate to radiation in the presence of a magnetic field and mark the thickness of the screen starting with which the traces of rays of each species vanish from the developed plate.

It turns out that the trace of  $\alpha$ -particles disappears first. Alpha-particles are completely absorbed by a sheet of paper of a thickness of about 0.1 mm (Figs. 377c and 378a). The flow of  $\beta$ -particles gradually attenuates with increasing thickness of the screen and is completely absorbed by an aluminium screen of a thickness of a few millimetres (Fig. 378b). The most

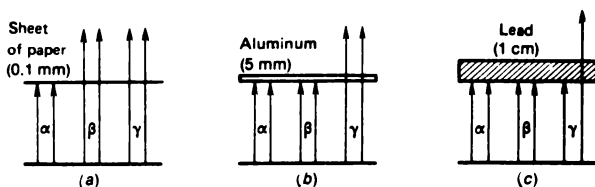


Fig. 378.

Absorption of radioactive radiation by substances.

penetrating is the  $\gamma$ -radiation. A 1-cm aluminium layer practically does not absorb the  $\gamma$ -radiation.

Substances having a large atomic number absorb  $\gamma$ -radiation stronger. In this respect,  $\gamma$ -radiation is similar to X-rays. For example, 1 cm of lead ( $Z = 82$ ) absorbs almost half a beam of  $\gamma$ -radiation (Fig. 378c).

The difference in the properties of  $\alpha$ -,  $\beta$ - and  $\gamma$ -radiations is visually manifested in the *Wilson cloud chamber*, viz. a device for observing the traces of fast charged particles. The Wilson chamber (Fig. 379) consists of a glass cylinder 1 with a glass lid, in which piston 2 can be moved. The volume of the cylinder under the piston is filled with air saturated with water (or alcohol) vapour. As the piston is rapidly moved downwards, the air in the chamber is cooled as a result of rapid expansion. Water vapour becomes supersaturated, i.e. the conditions for its condensation on condensation nuclei are created (see Vol. 1, Sec. 17.13). The products of air ionization may serve as condensation nuclei. Ions polarize water molecules and attract them, thus facilitating condensation. Dust particles may also serve as condensation nuclei, but air should thoroughly be purified before the Wilson chamber is used for observations.

Let us suppose that vapour in the chamber is supersaturated. A fast charged particle flying through the chamber leaves a chain of ions on its way. Since a water drop is formed on each ion, the trajectory of the particle can be seen in the form of a cloud track. Illuminating cloud tracks by a

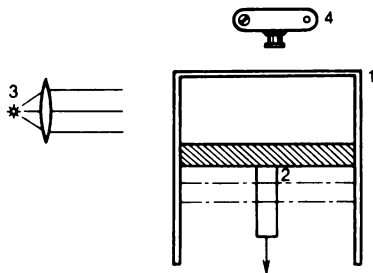


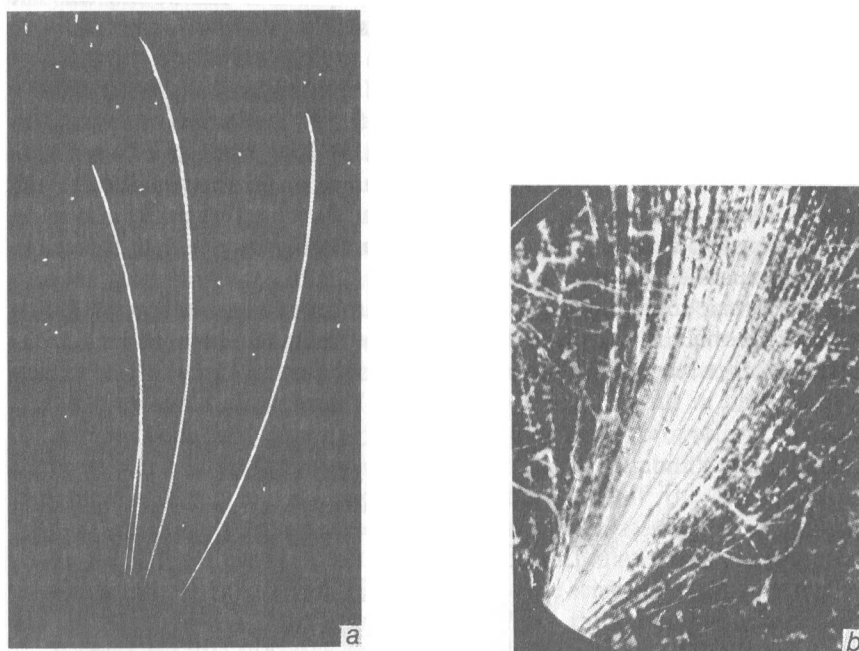
Fig. 379.

Wilson cloud chamber (simplified version): 1 — glass cylinder, 2 — piston, 3 — light source, 4 — camera. Air above the piston is saturated by water vapour.



powerful lamp 3 placed on one side of the chamber (see Fig. 379), we can photograph them through the transparent lid of the chamber. Such photographs are shown in Figs. 380 and 381. This wonderful method can be used to observe the trajectory (track) of a single  $\alpha$ - or  $\beta$ -particle. Cloud tracks exist in the chamber for a short time since air is heated from the chamber walls, and drops evaporate. In order to obtain new tracks, the available ions must be removed with the help of an electric field, and air must be compressed with a piston. After a certain time required for cooling the air in the chamber which has been heated as a result of compression, a new expansion can be carried out.

The value of the Wilson cloud chamber as a physical device considerably increases if it is placed in a magnetic field. This was done by the Soviet physicists P. L. Kapitza (1894-1984) and D. V. Skobeltsyn (b. 1892). The magnetic field bends the trajectories of particles (see Fig. 380). The direction of bending of a track makes it possible to judge about the sign of



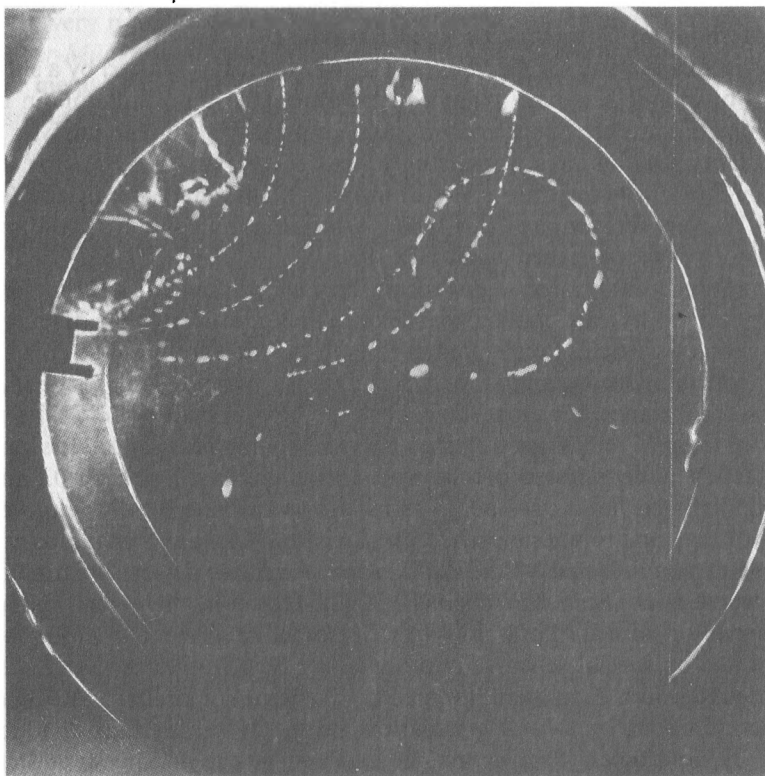
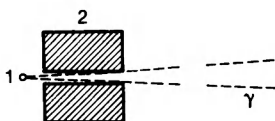
**Fig. 380.**

Traces of  $\alpha$ - and  $\beta$ -particles in a Wilson cloud chamber. The particles are emitted by a radioactive sample placed in the lower part of the chamber: (a)  $\alpha$ -particles: the chamber is in a magnetic field of 4.3 T, directed downwards perpendicular to the plane of the figure, (b)  $\beta$ -particles: a magnetic field of 0.0215 T is directed upwards at right angles to the plane of the figure.

the charge. Having measured the radius of the trajectory, one can determine the velocity of a particle if its mass and charge are known (see Sec. 22.5).

The length of tracks of  $\alpha$ -particles in air under the atmospheric pressure is about 5 cm and is much smaller than that of tracks of the majority of  $\beta$ -particles. The tracks of  $\alpha$ -particles are much denser than those of  $\beta$ -particles, which indicates a lower ionizing power of the latter.

Figure 381 shows a Wilson cloud chamber placed in a magnetic field and exposed to  $\gamma$ -radiation. The beams of  $\gamma$ -radiation are not deflected



**Fig. 381.**

The photograph of tracks in a Wilson cloud chamber placed in a magnetic field and exposed to  $\gamma$ -radiation. At the top, the location of the source: 1 — radioactive sample, 2 — lead screen with a slit, 3 —  $\gamma$ -radiation beam.

by the magnetic field, and their trajectories in the cloud chamber must be straight lines emanating from a source of  $\gamma$ -radiation. However, the photograph does not contain such straight tracks. Therefore,  $\gamma$ -radiation does not leave a continuous chain of ionized atoms on its path. The action of  $\gamma$ -radiation on a substance is reduced to occasional knockings out of electrons from atoms. The electrons which have received a high velocity at the expense of the energy of  $\gamma$ -quanta cause the ionization of atoms of the medium. The trajectories of these electrons bent by the magnetic field can be seen in Fig. 381. Most of electrons are emitted by the chamber walls.

It should be noted in conclusion that most of radioactive substances emit only one species of particles (either  $\alpha$ - or  $\beta$ -particles). The emission of particles is often (but not always) accompanied by the emission of  $\gamma$ -radiation.

### 23.3. Methods of Detecting Charged Particles

The devices capable to register a negligibly weak effect produced by a single particle of atomic size have played an exceptionally important role in the evolution of concepts of microcosm, and in particular, in the analysis of radioactivity. One of such wonderful devices is the Wilson cloud chamber which makes visible the trajectories of individual fast charged particles (see Sec. 23.2). Another device of this type is a scintillation counter whose simplified form was described in Sec. 22.10.

When some luminescent substances (like zinc sulphide or naphthalene) are bombarded by fast charged particles, a noticeable fraction of energy of the charged particles decelerated in the substances is found to be converted into visible light: an impact of a fast charged particle against the layer of such a substance causes a short flash of light known as *scintillation*. The brightness of scintillation flashes is especially high for  $\alpha$ -particles since an  $\alpha$ -particle is decelerated over a path whose length is less than 0.1 mm, and the liberated luminous energy turns out to be concentrated in a very small volume. Scintillations caused by particles on a screen made of zinc sulphide can be detected by the unaided eye. A simple device designed for this purpose is the *spinthariscopes* (Fig. 382). However, the visual method of observation of scintillations is very tiresome. At present, scintillations are counted by special sensitive photocells (see Sec. 21.4) known as *photoelectric multipliers*. They were invented by the Soviet physicist L. A. Kubetskii. Scintillations by  $\beta$ - and  $\gamma$ -radiation are much weaker than the glow caused by  $\alpha$ -particles. They cannot be detected by the naked eye and can be registered only with the help of photoelectric multipliers.<sup>2</sup>

<sup>2</sup> Detectors of particles consisting of layers of a luminescent substance and photoelectric multipliers can register single charged particles. These devices known as *scintillation counters* are widely used now. Modern experimental set-ups include hundreds or even thousands of such scintillation counters (see Chap. 24).

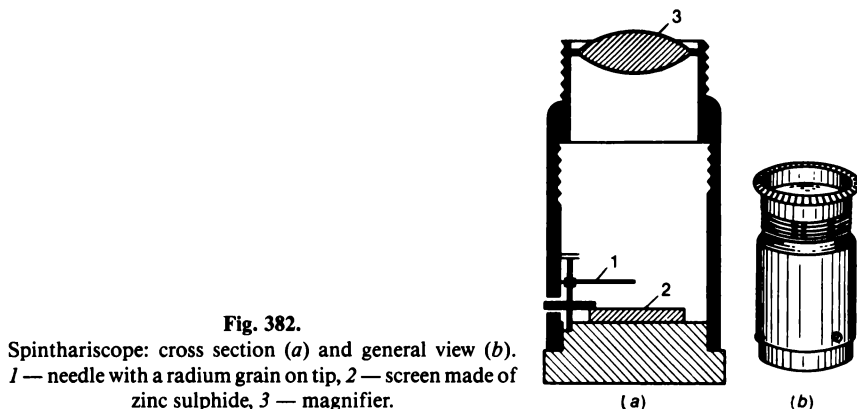


Fig. 382.

Spinhariscope: cross section (a) and general view (b).  
 1 — needle with a radium grain on tip, 2 — screen made of zinc sulphide, 3 — magnifier.

A very popular device for detecting individual charged particles is the *Geiger-Müller gas-discharge counter*. The gas-discharge counter (Fig. 383) is a metal cylinder 2 with a thin filament 1 stretched along its axis and insulated from the cylinder. The latter is filled with a gas mixture (say, argon + alcohol vapour) at a pressure of 100-200 mmHg. The filament is at a positive potential of about 1000 V relative to the cylinder.

The passage of any ionizing particle through the counter causes a short flash of gas discharge in it. As a result, a short current pulse passes through the circuit of the counter. If resistance  $R$  is large enough (about 1000 M $\Omega$ ), the potential of the filament is kept at a lower value for a few milliseconds, and this pulse can be detected from the deflection of the pointer of a sensitive electrometer 4. In actual practice, current pulses caused by the passage of a charged particle through the counter are amplified by a semiconductor or vacuum-tube amplifier and registered by the deflection of the pointer

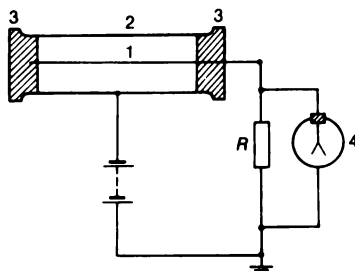


Fig. 383.

Gas-discharge counter: 1 — anode (thin filament), 2 — cathode (metal cylinder), 3 — insulators, 4 — electrometer for detecting discharges in the counter. During a discharge, electrons are accumulated at the filament of the counter, and its potential is lowered. After the discharge, the potential of the filament is restored due to the charges supplied by the battery through the resistor.

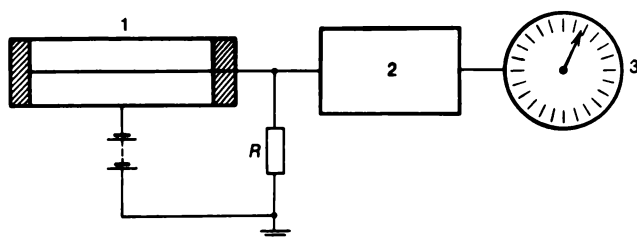


Fig. 384.

Schematic diagram of a device for registering radioactive radiations with the help of a gas-discharge counter: 1 — gas-discharge counter, 2 — amplifier, 3 — electromagnetic numbering machine,  $R \sim 1 \text{ M}\Omega$ .

of an electromagnetic annunciator or electromagnetic numbering machine connected to the amplifier (Fig. 384).

Let us consider in greater detail the mechanism of operation of a gas-discharge counter. The counter consists of two coaxial cylinders, and therefore the electric field in it is nonuniform (see Vol. 2, Sec. 2.19). The electric field strength attains the maximum value at the filament and rapidly attenuates with increasing distance from it (Fig. 385a). At a potential difference of about 1000 V, the electric field strength near the filament turns out to be high enough to impart to slow electrons a velocity required for ionizing the gas.

Let a free slow electron be formed somewhere in the counter (say, as a result of ionization of the gas under the action of a fast particle passing through the counter). This electron will move towards the positively charged filament and will ionize gas atoms in the region of the strong field near it. The electrons produced as a result of ionization are accelerated by the field and, in turn, cause ionization, giving rise to new and new electrons and causing new ionization.<sup>3</sup>

The number of ionized atoms increases in the avalanche-like manner, and an electric discharge emerges in the gas of the counter. The electrons formed in the discharge are very soon accumulated on the filament, while heavy and hence slow ions move towards the cylinder. The accumulation of electrons on the filament reduces its positive charge and gradually decreases the electric field strength at the filament (Fig. 385b). After a short time (of the order of a microsecond, i.e.  $10^{-6} \text{ s}$ ), the field is weakened to such an extent that it cannot impart to electrons a velocity required for ionization. The ionization ceases, then the discharge is discontinued.

If the counter is not connected to a battery, the electric field in it after the discharge remains weak, and a new discharge is impossible (Fig. 385c). But in practically used counters (see Figs. 383 and 384), the field in a counter is rapidly replenished due to the charge from the battery to which the counter is connected via resistor  $R$ . The counter turns out to be ready to operation again in  $100\text{--}200 \mu$  after the discharge.

It should be noted that the rapid quenching of a discharge occurs only with a special choice of the gas filling the counter and at a not very high voltage in it. For a too high voltage, a permanent discharge emerges in the counter, which consists of continuously repeated flashes of the type described above. The repetition of the discharge flashes is caused by the electrons knocked out from the cylinder of the counter by positive ions incident on it.

<sup>3</sup> In the electric field of the counter, positive ions acquire the same energy as that of electrons. However, we can neglect the ionization effect of positive ions since their velocity is very low due to a much larger mass.

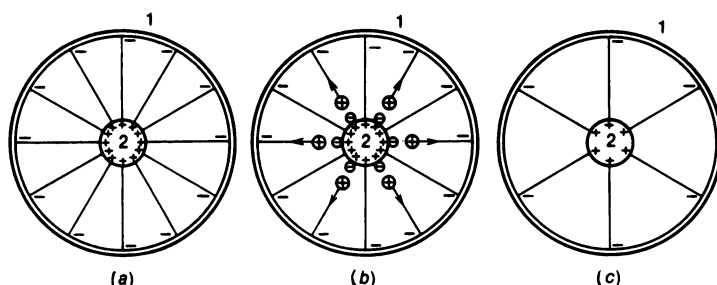


Fig. 385.

To the operation of a gas-discharge counter: 1 — cylinder of the counter, 2 — filament whose diameter is exaggerated. (a) The counter is charged to the working potential difference at which the passage of a charged particle through the counter causes a gas discharge in it. The electric field lines shown in the figure have the maximum density (electric field strength) at the filament. (b) The field of the counter at the moment when the discharge is self-quenched. The electrons formed during the ionization of the gas are accumulated at the filament and compensate a part of its positive charge. Positive ions continue to move to the cylinder. The field at the filament is weakened. (c) The field in the counter disconnected from the battery after the discharge has been quenched and the positive ions have reached the cylinder.

In the Geiger-Müller counter, the amplitude and duration of a current pulse emerging as a result of the avalanche process in the gas do not depend on the nature and energy of a charged particle being registered that “actuates” the counter (i.e. causes the avalanche process). One can select other operating conditions for a gas-discharge device and make it operate in the so-called proportional mode. If the voltage applied to the counter is reduced so that the avalanche process is not too intense to become a discharge, the number of ion pairs in such a “restricted avalanche” will be proportional to the initial degree of ionization. Such *proportional counters* not only register individual particles, but also measure the ionization caused by them (the energy losses of a particle in the gas), which is very important to identify particles.

Recently, the so-called semiconductor detectors have been designed. Such a detector is essentially an ionization chamber (see Fig. 376) in which air is replaced by a semiconductor. The utilization of specially prepared silicon or germanium makes it possible to reduce the dark current (viz. the current in the absence of an ionizing radiation) to the values admissible for recording an ionizing radiation. The advantage of semiconductor detectors is that they may absorb a large part of the energy of ionizing radiation due to the high density of a substance.

## 23.4. Properties of Radioactive Radiation

**1. Gamma-radiation.** In its properties,  $\gamma$ -radiation is similar to X-rays. Like X-rays, it ionizes air, blackens a photographic plate and is not deflected by a magnetic field. While passing through crystals,  $\gamma$ -radiation, like X-rays, undergoes diffraction. The two types of radiation are absorbed by screens the stronger, the larger the atomic number of the substance applied to a screen.

The penetrability of  $\gamma$ -radiation emitted by some radioactive substances is much higher than that of X-rays used in medicine and engineering.

However, the penetrability, or hardness, of X-rays increases with the voltage accelerating electrons. When the electrons accelerated by a voltage of about several million volts are decelerated, the formed X-rays have as high a penetrability as the hardest of radiations.

The coincidence of all properties of  $\gamma$ -radiation and hard X-rays points to their common nature. It is well known that X-radiation is a short-wave electromagnetic radiation. Consequently,  $\gamma$ -radiation is also formed by electromagnetic waves of a very short wavelength, and hence with a very high energy of quanta.<sup>4</sup> Like other types of electromagnetic radiation,  $\gamma$ -radiation propagates at a velocity of light equal to 300 000 km/s. Gamma-radiation and X-rays of the same wavelength do not differ from each other in any respect but the method of obtaining.

Measurements show that the energy of  $\gamma$ -quanta is different for different radioactive substances: there are  $\gamma$ -quanta with an energy of the order of ten kiloelectron volts (keV) to several megaelectron volts (MeV). The corresponding wavelengths range from  $10^{-10}$  to  $10^{-13}$  m.

**2. Alpha- and beta-particles.** In order to establish the nature of  $\alpha$ - and  $\beta$ -particles, it is necessary to measure the charge and the mass of an individual particle.

The measurement of charge does not present in principle any difficulty. We must measure the charge  $Q$  carried by a particle beam over a certain time and count the number  $N$  of particles passing during this time. The charge of a particle will obviously be

$$q = \frac{Q}{N}.$$

The charge of an  $\alpha$ - or  $\beta$ -particle can be measured experimentally as follows (Fig. 386a). A radioactive sample 1 emitting  $\alpha$ - or  $\beta$ -particles at a constant intensity is placed in front of diaphragm 2 whose aperture cuts a narrow particle beam. All particles passing through the aperture are trapped by a hollow metal cylinder 3 connected to a sensitive electrometer. The charge introduced by the beam into the cylinder is determined from the deflection of the electrometer.

Further, the electrometer and the cylinder are replaced by a particle counter 5 without changing the position of the sample and the diaphragm (Fig. 386b). The number of particles passing through the aperture of the diaphragm over the same time during which the charge has been measured is determined. The scintillation- or Geiger-Müller counters described in the previous section can be used in this experiment.

---

<sup>4</sup> It should be recalled that the energy  $W$  of a quantum is connected with the radiant frequency  $\nu$  and wavelength  $\lambda$  through the relations

$$W = h\nu = \frac{hc}{\lambda},$$

where  $h$  is Planck's constant and  $c$  is the speed of light in vacuum.

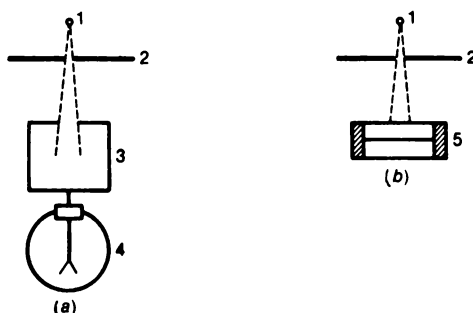


Fig. 386.

Schematic diagram of the experiment on measuring the charges of  $\alpha$ - and  $\beta$ -particles. (a) Measurement of the charge carried by a particle beam, (b) counting the number of passing particles. 1 — radioactive sample, 2 — diaphragm, 3 — collecting cylinder, 4 — electrometer, 5 — particle counter.

These experiments confirm that  $\alpha$ -particles carry a positive charge equal to two elementary charges. The charge of  $\beta$ -particles turned out to be a negative elementary charge.

The measurement of the masses of  $\alpha$ - and  $\beta$ -particles is a more complicated problem than the measurement of the mass of an ion (see Sec. 22.5) since the velocity of these particles is unknown. The experiments on the deflection of particles in a magnetic field do not allow us to determine both mass and velocity of a particle but give an idea about the ratio of these quantities. Another ratio of this kind can be obtained as a result of an additional experiment on the deflection of a particle in an electric field. Knowing two mass-to-velocity ratios for a particle, we can easily determine each quantity separately.

An experiment on simultaneous measurement of the mass and velocity of charged particles can be made as follows (Fig. 387). A beam of particles emitted by a radioactive source 1 gets into a narrow gap between the plates of capacitor 3 bent in the form of an arc of circle of radius  $\rho$ . Only those particles will pass through the gap containing an electric field

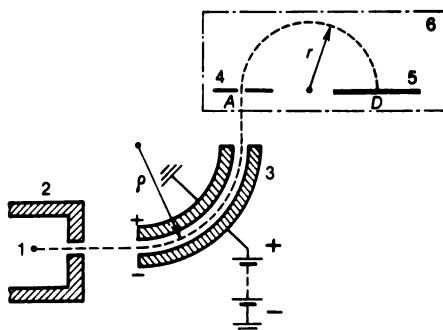


Fig. 387.

Schematic diagram of the simultaneous measurement of the velocity and mass of charged particles: 1 — radioactive source, 2 — screen with a slit, 3 — capacitor, 4 — diaphragm with a slit, 5 — photographic plate, 6 — magnet pole. The entire device is placed into an evacuated vessel which is not shown in the figure.



$E$ , whose masses and velocities are such that their trajectories in this field will be circles of radius  $q$ . The required centripetal acceleration  $v^2/q$  for these particles will be created by the electric force  $qE$ . Therefore,

$$\frac{mv^2}{q} = qE. \quad (23.4.1)$$

From the capacitor, the particles fly through slit 4 into a uniform magnetic field  $B$  whose lines are normal to the plane of the figure. Having described a semicircle in the magnetic field, the particles are incident on a photographic plate 5 at point  $D$ . After the development, the region of incidence of the particles is obtained in the form of a dark band. Measuring the length of segment  $AD$ , we determine the radius of the trajectory of a particle in the magnetic field. This radius  $r$  is connected with the velocity and mass of the particle through relation (22.5.1):

$$r = \frac{mv}{qB}.$$

Solving this equation and (23.4.1) for  $v$  and  $m$ , we can easily obtain

$$v = \frac{qE}{rB}, \quad m = \frac{qr^2B^2}{qE}.$$

The measurements of this kind lead to the following results. The mass of a  $\beta$ -particle coincides with the mass of an electron.<sup>5</sup> The charge of this particle is also the same as the electron charge. We hence arrive at the conclusion that  *$\beta$ -particles are electrons emitted by atoms of a radioactive substance*. The velocities of  $\beta$ -particles are huge and attain the values of 0.99 of the speed of light. Accordingly, the energy of  $\beta$ -particles may reach several megaelectron volts.

The mass of an  $\alpha$ -particle was found to be 4 amu. This mass and a charge of  $+2e$  correspond to the nucleus of a helium atom.

If an  $\alpha$ -particle is a helium nucleus, decelerated particles must combine with electrons to form helium atoms. This phenomenon was observed by Rutherford. He placed a radioactive sample into a glass ampoule with so thin walls that all  $\alpha$ -particles emitted by the sample emerged from the ampoule. The ampoule was placed into a thick-walled vessel of large volume. After a few days, the presence of helium in the outer vessel was detected.

Rutherford's experiment irrefutably proved that  *$\alpha$ -particles are fast helium nuclei*. The velocities of  $\alpha$ -particles are considerably lower than the velocities of  $\beta$ -particles and range from 10 000 to 20 000 km/s. The kinetic energy of an  $\alpha$ -particle is as high as 4-10 MeV.

---

<sup>5</sup> The mass of a particle depends on its velocity (see Sec. 22.6). Therefore, it would be more precise to formulate the result of the experiment on measuring the mass of a  $\beta$ -particle as follows: an electron and a  $\beta$ -particle of the same velocity have equal masses, or the rest masses of the electron and the  $\beta$ -particle are equal. In the case of  $\alpha$ -particles, this stipulation is immaterial since the velocity of an  $\alpha$ -particle is small as compared to the speed of light, and the mass measured in experiments is practically equal to the rest mass of the  $\alpha$ -particle.

As a result of collisions with the atoms of a medium, the energy of radioactive radiation is ultimately converted into heat. The thermal effect of radioactive radiation is readily revealed in calorimetric experiments.

### 23.5. Radioactive Decay and Radioactive Transformations

An analysis of radioactivity shows that *radioactive radiation is emitted by atomic nuclei of radioactive elements*. As regards  $\alpha$ -particles, this is obvious since they are not present in electron shells. The nuclear origin of  $\beta$ -particles can be proved in chemical experiments. If  $\beta$ -particles are emitted by nuclei, the  $\beta$ -radioactivity must cause a change in the chemical nature of an atom. Indeed, a  $\beta$ -electron carries away a unit negative charge from a nucleus, i.e. increases its positive charge by unity. The nucleus will keep  $Z + 1$  electrons near it instead of  $Z$  electrons, and the radioactive atom will be converted into an atom of the next element of the Periodic Table. Chemical analysis has proved indeed that atoms of an element with an atomic number exceeding by unity that of the atom emitting  $\beta$ -particles are accumulated in a substance emitting  $\beta$ -radiation.

The emission of  $\alpha$ -particles also changes the charge of the nucleus and therefore must also lead to a change in the chemical nature of a radioactive atom. This prediction is completely confirmed by experiments.

Thus, atoms of a radioactive element emitting  $\alpha$ - and  $\beta$ -particles change, being converted into atoms of a new element.

In this sense, the emission of radioactive radiation is termed as *radioactive decay*. There are two types of radioactive decay, viz. the  $\alpha$ -decay (emission of  $\alpha$ -particles) and the  $\beta$ -decay (emission of  $\beta$ -particles).<sup>6</sup>

Since an  $\alpha$ -particle carries away a positive charge of two units and a mass of four units, as a result of the  $\alpha$ -decay, a radioactive element is converted into another element with an atomic number two units less and a mass number<sup>7</sup> four units less. The mass of a  $\beta$ -particle is negligibly small in comparison with the atomic mass unit. Therefore, the emission of a  $\beta$ -particle does not change the mass number of the nucleus. Consequently, as a result of the  $\beta$ -decay, a radioactive element is converted into an element having the same mass number and an atomic number larger by unity.

These rules indicating the displacement of elements in the Periodic Table as a result of a radioactive decay are known as the *radioactive displacement law*.

---

<sup>6</sup> A detailed analysis of the  $\beta$ -decay revealed that along with  $\beta$ -particles, a very light neutral particle, viz. *neutrino* (see Sec. 25.2), is emitted. In the experiments illustrated by Fig. 377, a neutrino escapes detection since it produces neither ionization nor photographic effect.

<sup>7</sup> It should be recalled that the mass number of an atom or a nucleus is the atomic mass rounded to the nearest integer.

A radioactive decay causes a continuous decrease in the number of atoms of a radioactive element. For uranium, thorium and radium, the decay rate is so low that a decrease in the number of atoms of these elements is insignificant even over a time interval of several years. There is, however, a large number of rapidly decaying radioactive elements. Let us consider, for example, a  $\beta$ -radioactive isotope of bismuth with a mass number of 210, which is known as RaE (radium E). This substance is liberated from radium in which it is present in very small amounts. A negligibly small mass of RaE can easily be detected from an intense  $\beta$ -radiation. By measuring the number of  $\beta$ -particles emitted by a RaE sample per unit time periodically with the help of a gas-discharge counter, we find that this number gradually decreases. The curve illustrating the decrease in the activity<sup>8</sup> with time is shown in Fig. 388.

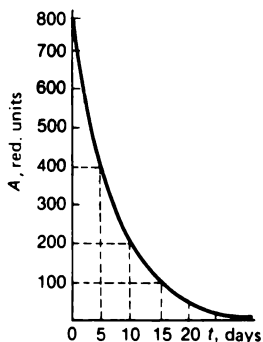
It can be seen from the figure that after five days the activity of RaE is equal to half the initial value, in 10 days, to 1/4 of the initial value, in 15 days, to 1/8 of the initial value, etc. Every five days the activity is reduced by half. However, in order to reduce the activity to half the initial value, it is sufficient to divide the sample into two parts. Consequently, the number of RaE atoms is reduced by half every five days.

The time interval over which half atoms of a radioactive substance decay is known as its *half-life*. Thus, the substance whose decay is represented in Fig. 388 has a half-life of 5 days. Let the number of atoms of the radioactive substance be  $N_0$  at the initial instant of time ( $t = 0$ ). We denote the half-life of this substance by  $T$ . After  $n$  half-lives, i.e. at the moment  $t = nT$ , the number of non-decaying atoms is obviously

$$N = N_0 \cdot 2^{-n}.$$

Substituting  $n = t/T$  into this expression, we obtain

$$N = N_0 \cdot 2^{-t/T}. \quad (23.5.1)$$



**Fig. 388.**  
Curve illustrating the decrease in activity  $A$  of the radioactive substance RaE with time.

<sup>8</sup> The *activity* of a radioactive sample is the number of particles emitted by it per unit time.

We derived this expression for the time period  $t$  multiple to the half-life (i.e. for integral  $n$ ). It can be proved, however, that it is valid for any  $t$  as well. Relation (23.5.1) describing the time dependence of the number of nondecaying radioactive atoms is called the *radioactive decay law*.

The half-life is one of the basic characteristics of a radioactive substance. Numerous experiments showed that the half-life of a radioactive substance is a strictly constant quantity which cannot be altered by any effect (in accessible limits) like cooling, heating, pressure, magnetic field, or chemical affinity. The independence of half-life of external conditions is not astonishing. Radioactive decay is a property of atomic nuclei which cannot be changed by using an energy of conventional macroscopic effects (see Sec. 22.15).

The measurement of the half-life of short-lived nuclei is reduced to determining the time interval during which the intensity of radiation is reduced by half. The half-life of long-lived nuclei can be calculated by measuring the number of atoms decaying per unit time (which is equal to the number of particles emitted during this time) and knowing the total number of atoms in a sample. Indeed, the fraction of atoms decaying during a certain time depends on the half-life. The shorter the half-life, the shorter time is required for the decay and the larger the fraction of atoms decaying during the same time.

The measurements of this kind result in the value of 1600 years for the half-life of radium. Naturally, the decrease in the amount of radium over time intervals of the order of a year is so small that the change in its activity is insignificant.

It is known from geology that the age of minerals is of the order of million years. Therefore, the decay of radium over time intervals significant on the geological scale would have led to its complete disappearance. Obviously, *along with radioactive decay, the formation of new radium atoms also takes place in nature*. The fact that radium is always present in uranium ores (and only in them) suggests that the radioactive decay of uranium serves as a source of new radium atoms.

Uranium is an  $\alpha$ -radioactive substance, i.e. it emits  $\alpha$ -particles. The half-life of uranium (to be more precise, its main isotope with an atomic mass of 238) measured from  $\alpha$ -activity amounts to 4.5 billion ( $4.5 \times 10^9$ ) years. This means that uranium decays very slowly even on the geological scale.

According to the radioactive displacement law, the  $\alpha$ -decay of the  ${}^{238}_{92}\text{U}$  nucleus<sup>9</sup> leads to the formation of a nucleus having a charge  $92 - 2 = 90$  and a mass number  $238 - 4 = 234$ , i.e. the thorium isotope  ${}^{234}_{90}\text{Th}$ . This

<sup>9</sup> It should be recalled that the numerals of the symbol of a chemical element have the following meaning: the subscript indicates the atomic number and the superscript is the mass number of an isotope under consideration.



Besides the uranium family, there are two more radioactive families. The parent of one family is thorium, while the parent of the other family is a rare uranium isotope  $^{235}_{92}\text{U}$ .

### 23.6. Applications of Radioactivity

**1. Biological effects.** Radioactive radiation is hazardous to living cells. The mechanism of this effect involves the ionization of atoms and the decomposition of molecules in the cells during the passage of fast charged particles. The cells in the state of fast growth and reproduction are especially sensitive to radioactive radiation. This is used for curing malignant tumours.

Radioactive samples emitting  $\gamma$ -radiation are used for therapeutic purposes since they penetrate living organisms without a noticeable attenuation. Moderate radiation doses kill malignant cells, without producing a significant damage to patient's organism. It should be noted that cancer radiotherapy, like X-ray therapy, is not a universal method leading to complete curing.

Too large radiation doses cause serious diseases in living organisms (radiation sickness) and may lead to death. On the contrary, small doses of radioactive radiation (especially  $\alpha$ -radiation) produce a stimulation effect on an organism. This explains the curative effect of radioactive mineral water containing small amounts of radium and radon.

**2. Luminescent compounds.** Luminescent substances glow under the effect of radioactive radiations (cf. Sec. 23.3). By adding a small amount of radium salt to a luminescent substance (say, zinc sulphide), luminous paints are obtained. Being applied to hands and dials of watches, sights, etc., these paints make them visible at dark.

**3. Determination of the age of the Earth.** The atomic mass of ordinary lead extracted from ores which do not contain radioactive elements is 207.2. It can be seen from Fig. 389 that the atomic mass of lead formed as a result of the uranium decay is 206. The atomic mass of lead contained in some uranium ores turns out to be very close to 206. It follows hence that these minerals did not contain lead at the moment of their formation (crystallization from melt or solution). The total amount of lead present in such minerals has been accumulated as a result of the uranium decay. Using the radioactive decay law, the age of a mineral can be determined from the ratio of the amounts of uranium and lead in it (see Ex. IV.32).

The age of minerals having different origin and containing uranium determined in this way is of the order of billion-years. The age of the most ancient minerals exceeds 1.5 billion years.

For the age of the Earth, we usually take the time elapsed from the moment of formation of its solid crust. The age of the Earth determined from a large number of measurements based on the radioactivity of uranium, thorium and potassium exceeds 4 billion years.

### 23.7. Accelerators

Fast  $\alpha$ -particle beams produced by radioactive samples were found to be an indispensable tool for probing atoms (see Sec. 22.10). But the role of fast particle beams in investigating atomic nuclei is probably even more significant (see Chap. 24). However, the investigation into nuclear structure requires charged particles of larger amounts and range and of higher velocities than those supplied by radioactive samples (i.e. not only  $\alpha$ -particles and electrons but also protons, deuterons<sup>10</sup> and nuclei of all chemical ele-

<sup>10</sup> Deuterons are nuclei of heavy hydrogen (deuterium).

ments). In order to satisfy this demand, various types of *accelerators*, viz. the devices for artificial acceleration of charged particles to high energies, have been developed.

The history of accelerators starts from 1932 when Rutherford's co-workers, the English nuclear physicist Sir John Douglas Cockcroft (1897-1967) and the Irish physicist Ernest Thomas Sinton Walton (b. 1903) constructed a set-up for obtaining protons with an energy of half-million electron volts. Since then, accelerator technology has been developed greatly, and modern accelerators can impart an energy of hundred billion electron volts to particles.

In order to impart an energy to a charged particle, it is sufficient to make it pass across an accelerating potential difference. By increasing this potential difference, we increase the energy of the particle. This, however, cannot be done unlimitedly because the electrical insulation may break down at a high voltage. The practical limit is a voltage of 5-8 MV. Such a voltage is obtained with the help of electrostatic generators (see Vol. 2, Sec. 2.20).

Since the potential difference cannot be increased beyond the limit of 8 MV, the energy of particles can be increased by accelerating them with the same potential difference *many times*. The idea of multiple acceleration of charged particles by a comparatively low potential difference forms the basis of the design of most modern accelerators. An example of the implementation of this idea is the *cyclotron* proposed in 1936 by the American physicist Ernest Orlando Lawrence (1901-1958). The operating principle of a cyclotron is as follows. Two hollow electrodes (known as *dees*) are mounted in a chamber continually evacuated to a very low pressure between the poles of a powerful magnet (Figs. 390-392). A rapidly varying potential difference is applied to the dees. A source of ions (for example, a gas dis-

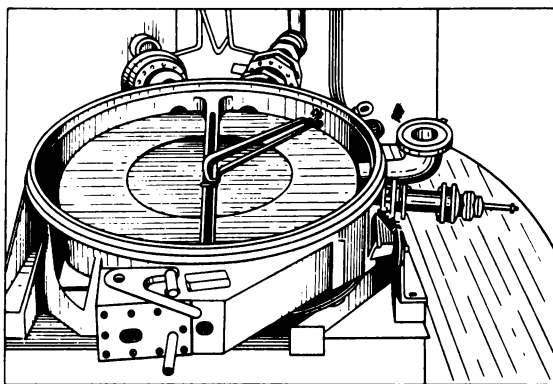
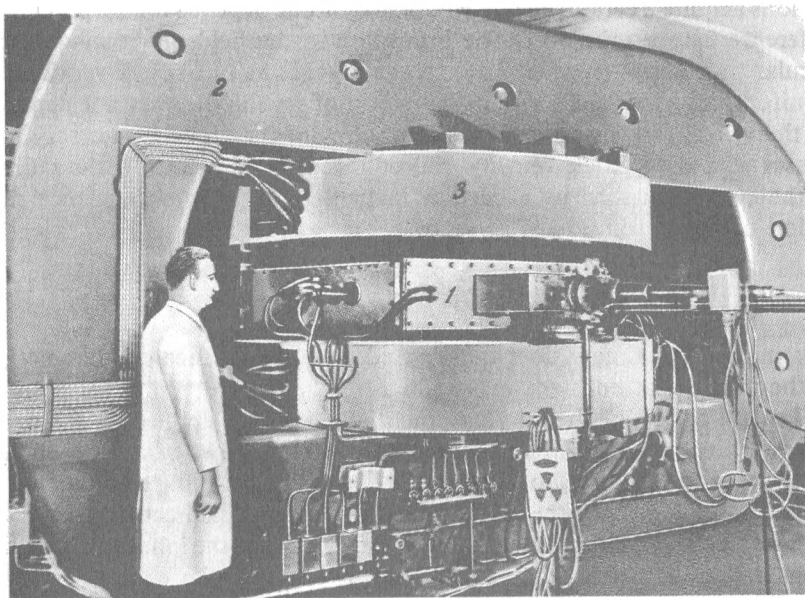


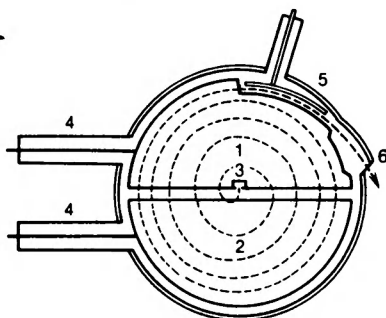
Fig. 390.

General view of the vacuum chamber of a cyclotron.

**Fig. 391.**

General view of a cyclotron accelerating protons to an energy of 25 MeV: 1 — vacuum chamber, 2 — yoke of an electromagnet, 3 — electromagnet poles with magnet windings.

charge in hydrogen) is mounted at the centre of the chamber between the dees (Fig. 392). During the half-periods of alternating current when the electric field is directed from dee 1 to dee 2, positive ions are pulled out from the slit of source 3. Having passed across the gap between the dees,

**Fig. 392.**

Schematic diagram of a cyclotron chamber: 1, 2 — dees (the dees resemble two halves of a very flat tin cut along the diameter), 3 — ion source, 4 — terminals of the alternating voltage applied to the dees, 5 — negatively charged plate for deflecting accelerated ions towards window 6 through which the accelerated ion beam is extracted. The dashed spiral indicates the trajectory of an ion.



the ions acquire a certain energy whose magnitude depends on the potential difference between the dees. The ions move in the field of the magnet in circular trajectories (see Sec. 22.5). A remarkable feature of the motion in a uniform magnetic field is that the period of revolution does not depend on the velocity of a particle since the radius of its circular trajectory increases with the particle velocity. Indeed, according to (22.5.1), the radius of the circle described by a particle in field  $B$  is  $r = mv/qB$ . Hence the period of one revolution is given by

$$\tau = \frac{2\pi r}{v} = \frac{2\pi}{v} \frac{mv}{qB} = \frac{2\pi m}{qB}.$$

This means that time  $\tau$  does not depend on  $v$ , and hence on the energy of the particle, at constant  $m$ ,  $q$  and  $B$ .

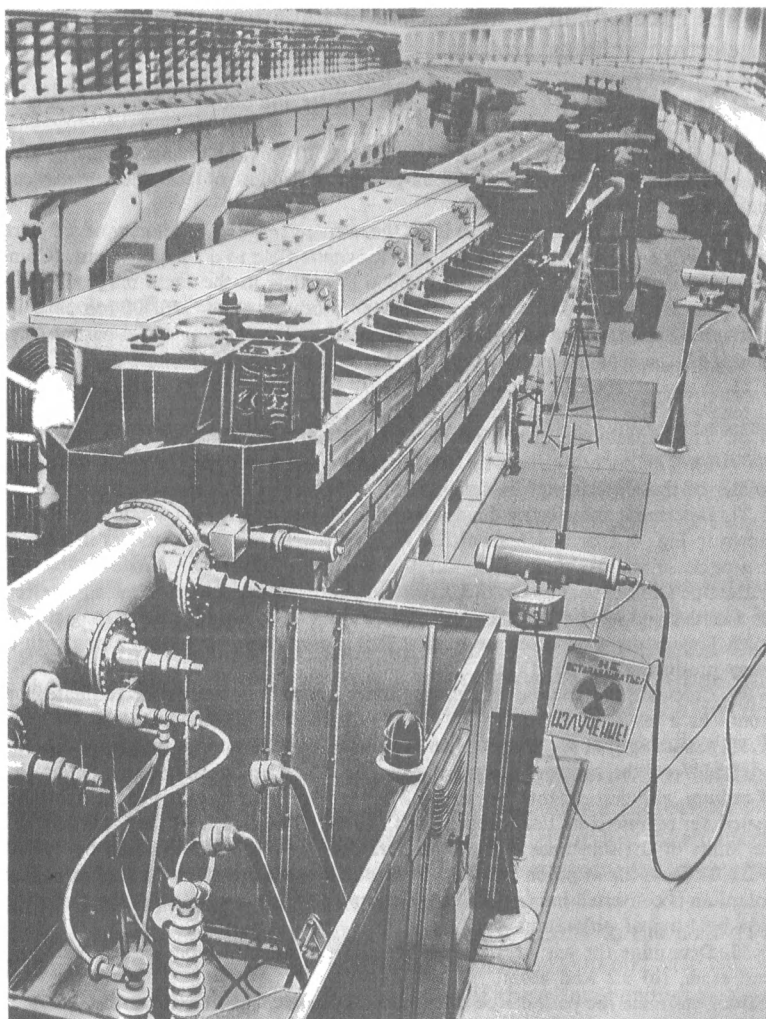
Let the period of the varying potential difference applied to the dees be equal to the period of revolution  $\tau$ . In this case, when an ion enters for the second time the gap between the dees having completed a half-turn in dee 2 (see Fig. 392), the electric field in the gap is directed from 2 to 1, i.e. its direction coincides with that of the motion of the ion. Consequently, having passed through the gap, the ion doubles its energy. Having described a half-turn in dee 1, the ion gets into the field in the gap, which is again directed from 1 to 2, and hence doubles its energy once more, and so on.

As the velocity of the ion grows, the radius of its trajectory increases according to (22.5.1). The trajectory of an ion in a cyclotron resembles an incoiling spiral. Using (22.5.1), we can easily calculate the energy of the ions which have been brought as a result of acceleration to a distance  $R$  from the source:

$$W = \frac{mv^2}{2} = \frac{q^2 b^2 R^2}{2m}.$$

The magnetic induction in a cyclotron cannot exceed 1.5 T because of the magnetic saturation of iron (see Vol. 2, Sec. 16.7). For this reason, in order to increase the energy of a particle, one has to increase the radius of the magnet poles. For example, the diameter of the magnet poles is 4.5 m in the cyclotron producing a proton beam with an energy of about 400 MeV.

When the kinetic energy of a particle being accelerated becomes comparable with its rest energy  $m_0 c^2$ , the dependence of the mass of the particle on its energy becomes noticeable. The mass of the particle increases with acceleration together with its period of revolution  $\tau$ . The latter becomes longer than the period of acceleration alternating voltage, and as a result the particle falls out of cycle and is no longer accelerated. It was believed some time ago that the increase in the mass with velocity sets a limit of



**Fig. 393.**

General view of a region of the circular hall of the Serpukhov 76-GeV accelerator: the diameter of the orbit is 472 m, the mass of iron in the magnet is 20 000 t.

100 MeV to the maximum energy of the particles accelerated in a cyclotron. The Soviet physicist V. I. Veksler (1907-1966) substantiated in 1944 a significant improvement of the operating principle of cyclotrons, which has allowed him to overcome the difficulty associated with the variable mass of particles. His discovery has made it possible to obtain particles accelerated

to energies of billion electron volts. The photograph in Fig. 393 shows a part of the circular hall of an accelerator imparting energies of about 76 billion electron volts to protons.

- ? **IV.22.** The capacitance of the filament of a counter and bodies connected to it is  $10 \text{ pF}$  ( $1 \text{ pF} = 10^{-12} \text{ F}$ ). What is the number of ion pairs formed during a discharge in the counter if the electrometer (see Fig. 383) indicates  $10 \text{ V}$ ? (The leakage of charge through a large resistance  $R$  during the discharge and deflection of the electrometer can be neglected.)
- IV.23.** The velocity of an  $\alpha$ -particle is on the average  $1/15$  of the velocity of a  $\beta$ -particle. Explain why  $\alpha$ -particles are deflected by a magnetic field to a smaller extent. (Compare the radii of the trajectories of an  $\alpha$ - and a  $\beta$ -particle in the same magnetic field.)
- IV.24.** An  $\alpha$ -particle with an energy of  $5 \text{ MeV}$  produces about  $150\,000$  ion pairs in air. Determine the ionization current created by a substance emitting  $100 \alpha$ -particles per second. (All the ions are accumulated at the electrodes.)
- IV.25.** Why are very high resistances  $R(10^8\text{-}10^{12} \Omega)$  required for measuring the ionization current with the help of an electrometer (see Fig. 368)?
- IV.26.** The capacitance of the electrometer in the set-up shown in Fig. 386 is  $10 \text{ pF}$  ( $10^{-11} \text{ F}$ ). The collector receives  $100\,000$  electrons per second. In what time will the pointer of the electrometer be deflected by a division if the value of division is  $0.1 \text{ V}$ ?
- IV.27.** Determine the electric field strength and the magnetic induction in the device shown in Fig. 387 for which a particle with an energy of  $100 \text{ keV}$  moves in both fields in a circle of radius  $20 \text{ cm}$ . Consider the case of  $\alpha$ - and  $\beta$ -particles.
- IV.28.** One gram of radium emits  $3.7 \times 10^{10}$   $\alpha$ -particles per second. How many electrons are emitted per second by RaE accumulated in  $1 \text{ g}$  of radium over a long time?
- IV.29.** How many  $\alpha$ -particles are emitted by  $1 \text{ g}$  of radium per second along with other decay products? (see Ex. IV.27.)
- IV.30.** Calculate the volume of helium (under normal conditions) accumulated over a month as a result of the decay of  $1 \text{ g}$  of radium with descendants.
- IV.31.** Assuming that the energy of an  $\alpha$ -particle is  $5 \text{ MeV}$  and the energy of a  $\beta$ -particle is  $0.5 \text{ MeV}$  on the average, determine the amount of heat liberated per minute by  $1 \text{ g}$  of radium with descendants (see Ex. IV.27). By what temperature will the substance be heated per minute if its heat capacity is  $4.18 \text{ J/K}$ ? (The escape of the  $\beta$ -particle beyond the limits of the substance and the heat transfer from the latter should be neglected.)
- IV.32.** The half-life of polonium  $^{210}\text{Po}$  is  $140$  days. As a result of emission of an  $\alpha$ -particle, polonium is converted into stable lead. Determine the amount of lead liberated over  $100$  days by  $1 \text{ mg}$  of polonium.
- IV.33.** Determine the age of a mineral in which a uranium atom corresponds to (a) a lead atom, (b)  $0.2$  lead atom.
- IV.34.** Determine the period  $\tau$  of the alternating voltage accelerating protons in a cyclotron. The magnetic induction of the field is  $1.5 \text{ T}$ .
- IV.35.** Protons are ejected from a cyclotron as soon as they attain radius  $R = 100 \text{ cm}$ . What is their energy if  $B = 1.5 \text{ T}$ ? How many turns are completed per unit time by a proton emitted by an ion source by the moment when the potential difference between the dees is  $100 \text{ kV}$ ?

## Chapter 24

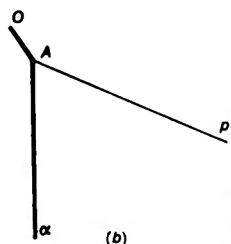
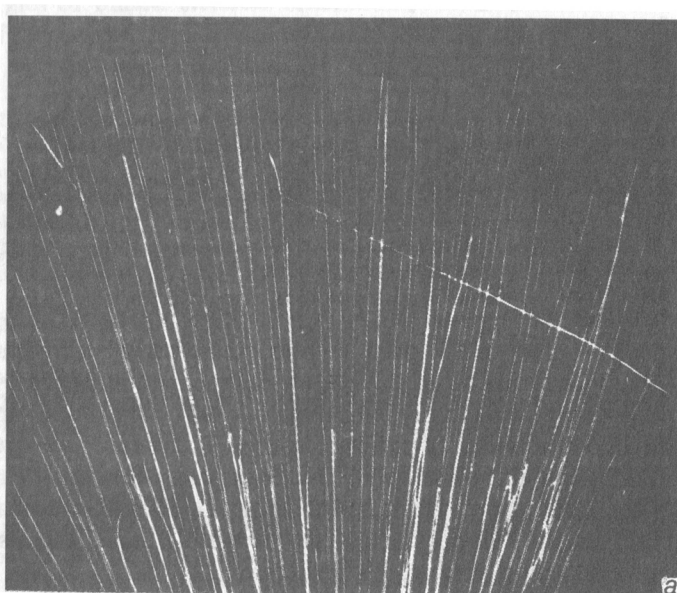
# Atomic Nuclei and Nuclear Power

### 24.1. Nuclear Reactions

In the previous chapter, we discussed the phenomena associated with radioactivity. It was shown that atomic nuclei of radioactive elements are unstable, i.e. they decay with time, emitting  $\alpha$ - and  $\beta$ -particles, and are transformed into the nuclei of other elements. These facts indicate that in spite of their negligible size, *atomic nuclei are compound particles consisting of other simpler particles*. It was mentioned earlier that radioactivity not only points to the composite structure of atomic nuclei but also provides the tools for investigating this structure.

One of such tools are fast  $\alpha$ -particles which can penetrate light nuclei and split them into components. Rutherford was the first to observe in 1919 the splitting of an atomic nucleus under the action of  $\alpha$ -particles. Continuing the experiments described in Sec. 22.10, he noted that when particles bombard nitrogen, boron and other elements, new particles are created, which also produce scintillations, but differ from  $\alpha$ -particles in higher penetrability. With the help of magnetic deflection and other methods, he managed to determine the charge and mass, and hence the origin of these particles. It turned out that they were fast nuclei of hydrogen atoms. (It should be recalled that the nucleus of a hydrogen atom, viz. a *proton*, has a mass very close to 1 amu and a charge  $+e$ .)

The process of proton emission was investigated with the help of a Wilson cloud chamber. An  $\alpha$ -radioactive substance was placed into the Wilson chamber filled with nitrogen. The volume of the chamber was periodically increased, and the obtained pattern was photographed. On the photograph, the tracks of  $\alpha$ -particles emerging from the substance could be seen (Fig. 394a). Most of tracks were linear. In some cases, however, the track of an  $\alpha$ -particle formed a “fork” at a certain distance from the end of the track (see diagram in Fig. 394b) viz. bifurcated into two different tracks of which the longer (*p*) was thinner and the shorter (*O*) was thicker than the track of the  $\alpha$ -particle. The formation of such a “fork” could be explained only as the result of collision of the  $\alpha$ -particle with the nucleus of a nitrogen atom.

**Fig. 394.**

Splitting of a nitrogen nucleus by an  $\alpha$ -particle in a Wilson cloud chamber: (a) the photograph of tracks in the chamber, (b) schematic diagram of tracks forming a "fork"; the track of the  $\alpha$ -particle colliding at point *A* with the nitrogen nucleus is marked by  $\alpha$ , while *p* and *O* indicate the products of splitting (a proton and an oxygen nucleus).

On the basis of Rutherford's observations, we must ascribe one of the tracks of the "fork" to a proton. Because of its smaller charge, the proton acts on the electrons of a nitrogen atom with a smaller force in comparison with that exerted by the  $\alpha$ -particle. Therefore, the proton causes smaller ionization per unit path length and forms the thinner track in the Wilson cloud chamber. The thicker track belongs to a particle whose ionization effect is stronger than that of the  $\alpha$ -particle, and which therefore has a larger charge.

The origin of this particle can be established by using the charge and mass conservation laws. Before the collision, we had two particles: (1) the  $\alpha$ -particle (i.e. the nucleus of a helium atom) having a charge of +2 units and a mass of 4 units and (2) the nucleus of a nitrogen atom having a charge of +7 units and a mass of 14 units. The total charge is +9 units, while the total mass is 18 units. After collision, we also have two particles

one of which is a proton (i.e. the nucleus of a hydrogen atom) having a charge +1 and a unit mass. The other particle must have a charge of +8 and a mass of 17.

The eighth element of the Periodic Table is oxygen. Thus, the "fork" indicates that we have a conversion of nitrogen and helium nuclei into the nuclei of oxygen and hydrogen respectively.

Following Rutherford's discovery, some other processes of this kind were registered, in which the nuclei (and hence the atoms) of some elements are transformed into the nuclei (atoms) of other elements. Such processes are called *nuclear reactions*.

The nuclear reaction *nitrogen + helium = oxygen + hydrogen* is symbolically written as follows:



In this notation, the superscripts express the law of mass conservation ( $14 + 4 = 17 + 1$ ), while the subscripts express the law of charge conservation ( $7 + 2 = 8 + 1$ ).

## 24.2. Nuclear Reactions and Transformation of Elements

The study of nuclear reactions was stimulated by the development of devices for imparting high energies to charged particles, i.e. *accelerators* (see Sec. 23.7). Accelerators produce intense beams of charged particles accelerated to energies which can exceed many times the energy of particles of radioactive radiation.

It turned out that artificially accelerated fast protons, deuterons (nuclei of heavy hydrogen), helium nuclei and nuclei of other, heavier elements are capable to cause various splittings similar to the reaction of  $\alpha$ -particles (emitted by a radioactive substance) with nitrogen, which was considered by us earlier.<sup>1</sup> Fast electrons and X-ray photons (obtained by decelerating electrons that have been accelerated to 10 MeV and higher energies) also cause nuclear reactions. However, the effectiveness of splitting by photons, and especially by electrons, is inferior to that caused by heavy particles (protons, deuterons and other accelerated nuclei).

A large number of various nuclear reactions are known at present. Some of them will be considered later.

Of great importance was the discovery of neutral (uncharged) particles among the products of nuclear reaction. These particles have a mass equal to the mass of a proton (i.e. about 1 amu). The properties of these particles called *neutrons* will be discussed in the next section.

<sup>1</sup> Nuclear reactions may be caused by nuclei heavier than  $\alpha$ -particles if their energy is high enough for overcoming electrostatic repulsion. At present, nuclear reactions caused by relativistic heavy ions attract the attention of researchers.

The discovery of nuclear reactions was important in principle since the possibility of artificial transformation of elements has been proved for the first time. True, at the early stage only insignificant amounts of a substance could be transformed. This is so since the number of fast particles produced by accelerators or radioactive substances is relatively small (see Ex. IV.38 at the end of the chapter). Besides, only a small fraction of these particles initiates nuclear reactions. For example, from  $10^5$ - $10^6$  bombarding  $\alpha$ -particles, only one causes a nuclear transformation.

Such a small number of nuclear reactions can easily be explained. In order to penetrate an atomic nucleus, an incident charged particle must overcome huge forces of electrostatic repulsion since both the particle and the bombarded nucleus have positive charges. For this reason, nuclear reactions can be caused only by sufficiently fast particles capable of overcoming the electrostatic repulsion. However, fast particles moving in a substance spend their energy to ionize and excite atoms. Very soon they are decelerated completely, capture electrons and become neutral atoms. On account of the very small size of nuclei (see Sec. 22.10), only a few particles collide with a nucleus before they have lost their energy. Only such rare cases result in nuclear splitting.

It will be shown in Sec. 24.9 that under the conditions existing on stars, nuclear reactions, having been initiated once, proceed spontaneously like a fire consuming new and new portions of fuel burns as long as the fuel is available. Such self-sustained or nonstopped *chain reactions* can also be realized in terrestrial conditions (see Sec. 24.10).

In chain reactions, transformations of atoms occur on the large scale often comparable with that on which molecules are transformed in chemical reactions.

### 24.3. Properties of Neutrons

Emission of neutrons was detected for the first time in 1932 during the bombardment of beryllium by  $\alpha$ -particles. As a result of the capture of an  $\alpha$ -particle by a beryllium nucleus, a carbon nucleus is formed and a neutron is emitted. The reaction equation can be written as follows:



Here the symbol  $n$  denotes a neutron. Later, a large number of nuclear reactions involving the liberation of neutrons have been discovered. However, the reaction of  $\alpha$ -particles with beryllium is used, as before, to obtain neutrons. Even now ampoules filled with a mixture of an  $\alpha$ -radioactive substance and beryllium powder are used as compact neutron sources.

Let us place such a neutron source near an operating Wilson cloud

chamber in which a film made of a hydrogen containing substance (say, paraffin  $C_{22}H_{46}$ ) is fixed. On a photograph taken in the chamber, we shall see tracks emerging from the film. These are proton tracks as can be judged from the nature of ionization (see schematic diagram in Fig. 394). All the tracks are in the forward direction (if viewed from the source). They are produced by protons knocked from the film as a result of impacts of fast neutrons emitted by the source. There are no tracks of neutrons themselves which cross the chamber. Consequently, neutrons do not produce a noticeable ionization, i.e. unlike charged particles, they practically do not interact with electrons. Passing through a substance, neutrons interact only with atomic nuclei, but since the size of nuclei is very small, neutrons collide with them only rarely (see Ex. IV.40 at the end of the chapter). This explains the ability of neutrons to freely pass through thick (of the order of a centimetre) layers of substances (for example, the walls of the source and the chamber in the experiment shown in Fig. 395).

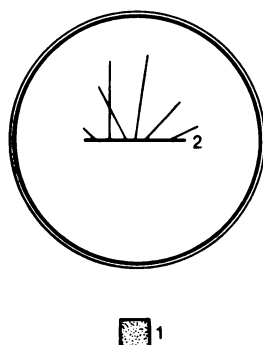


Fig. 395.

Schematic diagram of a Wilson cloud chamber irradiated by neutrons: 1 — source of neutrons (ampoule containing a mixture of an  $\alpha$ -radioactive substance and beryllium), 2 — paraffin film. The neutrons knock out fast protons of the paraffin film, which leave tracks in the chamber.

It can be seen from Fig. 395 that protons knocked out in the direction of motion of neutrons have longer tracks (and hence the higher energy) than those at a noticeable angle to this direction. This can easily be explained by analyzing the collision of a neutron with a proton as a collision of elastic balls having equal masses. The ball flies in the exactly forward direction only as a result of a head-on collision with the other ball (Fig. 396). But in this case the impinging ball comes to a halt, i.e. imparts its entire energy to the ball at rest. The motion of a proton at an angle to the initial velocity of a neutron corresponds to an oblique impact (Fig. 397). In this case, the impinging ball does not come to a halt but changes its direction of motion, having imparted only a part of its energy to the second ball.

During a collision of balls having different masses, the energy transfer is not so significant as during the collision of balls of the same mass, the amount of transferred energy being the smaller, the larger the difference



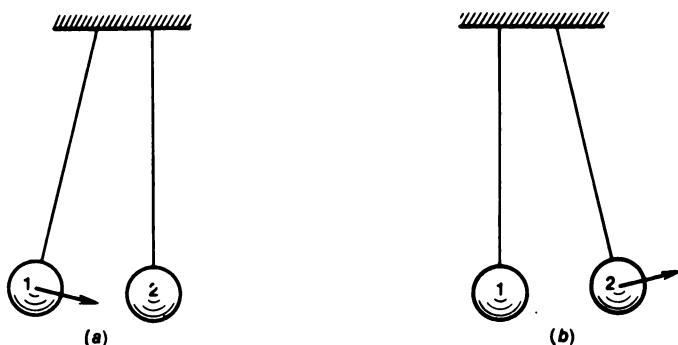


Fig. 396.

The head-on collision of elastic balls of equal masses: (a) before the collision, (b) after the collision. The impinging ball comes to a halt, imparting its velocity to the ball at rest.

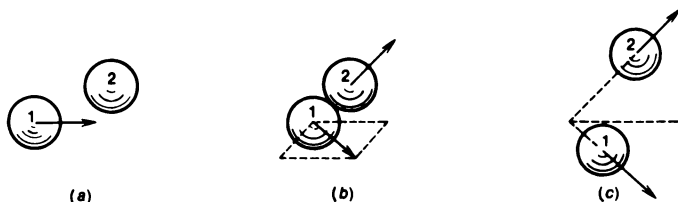


Fig. 397.

Glancing collision of elastic balls of equal masses: (a) before the collision, (b) during the collision, (c) after the collision. The ball that was initially at rest flies at an angle to the initial direction of motion of the impinging ball and receives only a fraction of its energy.

in the masses of the balls. Indeed, a light ball impinging at a heavy ball is bounced back, retaining almost the whole of its energy. Only a small fraction of the energy of the light ball is transferred to the heavy ball (see Ex. IV.41).

The analogy with the collision of balls leads to the following conclusion. *During collisions with nuclei, neutrons lose their energy, i.e. are decelerated.* The decelerating effect of collisions is the stronger, the lighter a nucleus, i.e. the closer the mass of the nucleus to the mass of the neutron. An especially strong deceleration takes place when neutrons collide with protons having the same mass as that of neutrons.

#### 24.4. Nuclear Reactions Induced by Neutrons

A collision of a fast neutron with a nucleus mainly leads to the scattering of the neutron, i.e. the change in the direction of its flight and the transfer of a fraction of its energy to the nucleus. However, another result of collision is also possible: a neutron can be captured by a nucleus, and a nuclear reaction may occur. An example of the neutron-induced nuclear reaction

is the splitting of boron:



Having captured a neutron, a boron nucleus splits into the nuclei of lithium and helium flying apart at a high velocity.

A reaction of boron with neutrons can be observed by placing a thin layer of boron into a Wilson cloud chamber. Irradiating the chamber by fast neutrons, we shall see on the photographs thick tracks of lithium and helium nuclei emanating from the boron layer in all directions (Fig. 398a).

Let us surround a source of neutrons with a substance containing a large amount of hydrogen, for example, with a paraffin sphere 15-20 cm in diameter. On their way to the chamber, neutrons will collide with carbon atoms ( $A = 12$ ) and, which is especially important, with protons. As was shown in the previous section, neutrons will be decelerated in this process and will get into the Wilson cloud chamber at an energy which is much smaller than their initial energy.<sup>2</sup> The effect of paraffin will be unexpected: the number of tracks on the photographs, and hence the number of splitting acts with boron nuclei, will be many times larger (Fig. 398b). Consequently, *the slower the neutrons, the more effective is their capture by the nuclei and the initiation of nuclear reactions.*

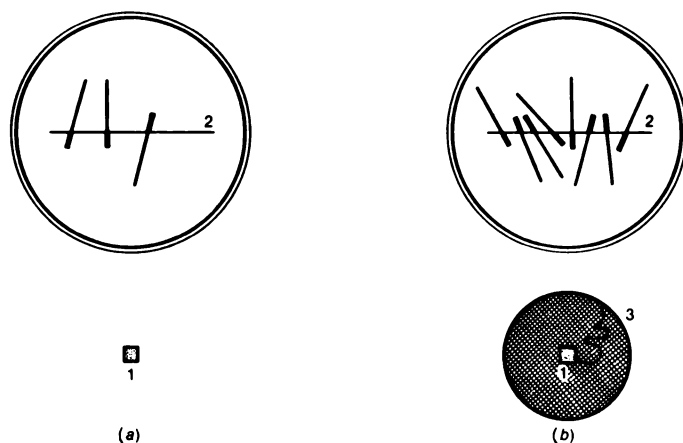


Fig. 398.

Schematic diagram of the experiment for observing the splitting of boron by fast (a) and slow (b) neutrons: 1 — source of neutrons, 2 — thin boron film in a Wilson cloud chamber, 3 — paraffin sphere. Short dense tracks are due to  ${}^7_3\text{Li}$  nuclei, while longer tracks are left by  $\alpha$ -particles. The dashed line shows the track of a neutron in the paraffin sphere.

<sup>2</sup> As a rule, neutrons are emitted by sources at an energy above 1 MeV. Moving in a thick layer of paraffin, they are decelerated to such an extent that their energy is lowered to the energy of thermal motion of atoms of the medium (0.03-0.04 eV).

Besides the neutron velocity, the effectiveness of their capture by a substance depends on the nature of the atoms. Analyzing the passage of slow neutrons through a boron layer, we shall find that they are absorbed almost completely by a boron layer whose thickness is fractions of a millimetre. Similar experiments indicate that besides boron, cadmium, lithium, chlorine and silver are also strong absorbers of slow neutrons. On the contrary, such substances as beryllium, deuterium, carbon and bismuth absorb slow neutrons to a very small extent.

The strong absorption of slow neutrons by nuclei is due to the absence of electrostatic repulsion between them (since a neutron does not carry any electric charge), and due to the attraction between nuclei and neutrons (see Sec. 24.8). A fast neutron flies by a nucleus during a very short time interval so that attractive forces have no time to deflect it and pull into the nucleus. The slower the motion of a neutron, the longer the time during which it experiences the attraction by the nucleus, and the easier it is captured by the nucleus. The capture of neutrons by nuclei is a reason for which neutrons do not exist in free state for a long time. The other reason is the radioactivity of neutrons. Experiments show that after a certain time a free neutron becomes a proton as a result of emission of an electron and a neutrino (see Sec. 25.1). The neutron half-life is about 15 minutes.

### 24.5. Artificial Radioactivity

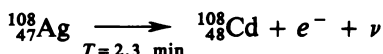
Investigating nuclear decay reactions, the French physicists F. Joliot-Curie (1900-1958) and Irene Curie (1897-1955) discovered in 1934 that decay products are often radioactive. Radioactive substances formed as a result of nuclear reactions were called *artificial radioactive substances* in contrast to *naturally radioactive substances* occurring in minerals (see Sec. 23.1).

Artificial radioactive substances can be obtained in various nuclear reactions. An example is the capture of neutrons by silver. In order to carry out this reaction, it is sufficient to place a silver plate near a neutron source enclosed in paraffin. As was mentioned in Sec. 24.4, neutrons are decelerated in paraffin, and slow neutrons are readily captured by nuclei and cause nuclear reactions. The silver plate does not exhibit any visible change under the action of neutrons. However, one can easily verify that some changes have occurred if the silver plate that has been bombarded by slow neutrons during a few minutes is brought to a gas-discharge counter. The counter will indicate that the plate has become radioactive, i.e. it emits radiation registered by the counter. It can be verified that the plate emits electrons ( $\beta$ -radiation). It turns out that the radioactivity acquired by silver gradually attenuates, being reduced by half in 2.3 min. Thus, a radioactive substance with a half-life of 2.3 min has been formed in ordinary silver. Auxiliary experiments as well as theoretical considerations indicate that the nuclear

reaction occurs according to the scheme



(letter  $\gamma$  on the right-hand side of this formula shows that  $\gamma$ -rays are emitted in this reaction). The atoms of the silver isotope  ${}^{108}\text{Ag}$  were found to be  $\beta$ -radioactive. They decay, emitting electrons and neutrinos (symbol  $\nu$ )<sup>3</sup>, and are converted into the atoms of a stable cadmium isotope:

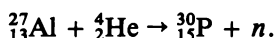


(the notation under the arrow indicates that the half-life is equal to 2.3 min). The radioactivity of the  ${}^{108}\text{Ag}$  isotope explains why the isotope with a mass number of 108 does not occur in natural silver which is a mixture of isotopes with mass numbers of 107 and 109: this isotope has a short lifetime and decays almost completely after its formation.

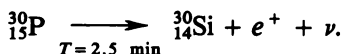
Artificial radioactivity is a common phenomenon. At present, several artificial radioactive isotopes for each element of the Periodic Table have been obtained. The total number of known artificial radioactive isotopes exceeds 1500, while the number of natural radioactive isotopes is 40, the number of stable (nonradioactive) isotopes being 260.

All the three types of radiation ( $\alpha$ ,  $\beta$  and  $\gamma$ ) characterizing natural radioactivity are also emitted by artificial radioactive substances.<sup>4</sup> However, artificial radioactive substances exhibit one type of decay which is not observed for natural radioactive substances. This is the decay involving the emission of *positrons*, viz. particles having the same mass as electrons and carrying a positive charge. The charges of an electron and a positron are equal in magnitude.

By way of an example of obtaining a positron-active substance, let us consider the reaction discovered by Joliot-Curie:



As a result of bombardment of aluminium by  $\alpha$ -particles, a neutron is emitted and a phosphorus isotope with a mass number of 30 is formed. Natural phosphorus contains only one isotope with a mass number of 31. The isotope  ${}_{13}^{30}\text{P}$  formed in the above reaction is radioactive and decays, emitting positrons (denoted by  $e^{+}$ ) and neutrinos according to the scheme



The half-life of  ${}_{13}^{30}\text{P}$  is 2.5 min. Its decay product is the stable silicon isotope  ${}_{14}^{30}\text{Si}$ .

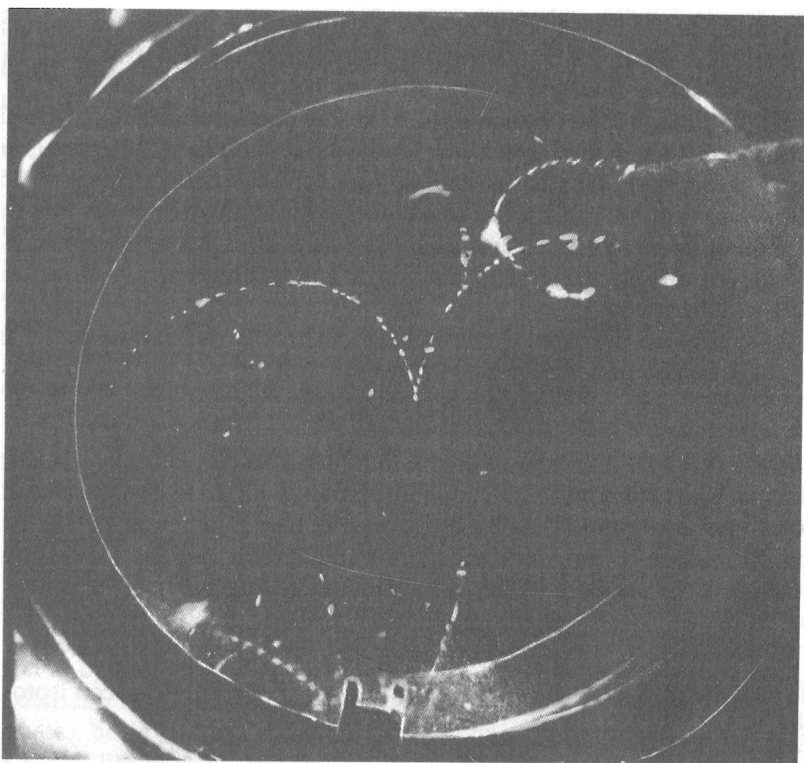
<sup>3</sup> See footnote to p. 455.

<sup>4</sup> Emission of  $\alpha$ -particles is observed only for elements at the end of the Periodic Table.

### 24.6. Positron

The first indications of the existence of positrons, viz. light particles differing from electrons only in the sign of charge, were obtained in 1932 with the help of a Wilson cloud chamber. A thin track left by an obviously singly charged and very light particle resembling the electron but bent in the direction corresponding to a positive charge was observed in the Wilson chamber placed in a magnetic field (see Ex. IV.45 and Fig. 413). Later it was established that two main processes responsible for the formation of positrons are artificial radioactivity and its interaction of high-energy  $\gamma$ -quanta with atomic nuclei. The former of these processes was considered in the previous section.

The latter process can be investigated with the help of a Wilson cloud chamber placed in a magnetic field and irradiated by a narrow beam of  $\gamma$ -radiation. In some photographs, peculiar twin tracks can be seen on the path of the beam of  $\gamma$ -radiation. One such track is shown in Fig. 399.



**Fig. 399.**

An electron-positron pair formed by a  $\gamma$ -quantum in a Wilson cloud chamber. The left track belongs to the electron and the right track, to the positron.

A charged particle moving in a gas loses energy to ionize gas atoms, and hence its velocity continuously decreases. But the lower the velocity of a particle, the stronger its trajectory is bent by the magnetic field (see Sec. 22.5). A thorough examination of the track reveals that the bend of each of its branch becomes sharper with increasing distance from the kink of the track. This means that we deal with the tracks of a pair of particles emerging from a point rather than with a curved track of a single particle. Judging by the degree of ionization, the two tracks are similar to the tracks of electrons. They are bent by the magnetic field in opposite directions, i.e. belong to oppositely charged particles. The information presented earlier makes it possible to conclude that one of the particles is an electron, while the other is positron.

Thus,  $\gamma$ -quanta passing through a substance (gas in the Wilson cloud chamber) form positrons in pairs with electrons rather than single particles. This phenomenon became known as the *formation of electron-positron pairs*. The theory indicates that a pair is formed as a result of interaction of a  $\gamma$ -quantum with the electric field of an atomic nucleus of a substance. In this process, the  $\gamma$ -quantum is converted into an electron-positron pair and the nucleus remains unchanged.

Using radioactive substances as intense sources of positrons, it has become possible to study the properties of these substances in detail. It has been proved, in particular, that the mass of the positron is exactly equal to the mass of the electron, i.e. constitutes about  $1/2000$  of the mass of the proton.

A process inverse to the formation of electron-positron pair has also been discovered. It turned out that an electron and a positron having approached each other to a short distance under the action of forces of electrostatic attraction may form two  $\gamma$ -quanta flying apart in the opposite directions (Fig. 400). The process of combining an electron and a positron accompanied by their conversion into  $\gamma$ -quanta is known as annihilation<sup>5</sup>.

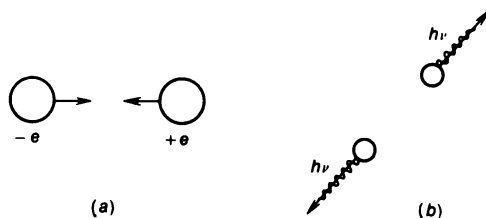


Fig. 400.

The combining of a positron and an electron, which generates a  $\gamma$ -quantum: (a) the electron and the positron attract each other, (b) as a result of collision, the electron and the positron are converted into two  $\gamma$ -quanta.

<sup>5</sup> Annihilation is not a very suitable term since it is derived from the Latin *nihil* meaning nothing.

Annihilation is the reason behind the absence of positrons on the Earth. After a negligibly short time following its formation, each positron is combined with an electron of the medium and is converted into two  $\gamma$ -quanta.

The formation of electron-positron pairs from  $\gamma$ -quanta and the combining of electrons with positrons leading to the formation of two  $\gamma$ -quanta are basically new processes in which the *mutual conversion of electromagnetic field of radiation ( $\gamma$ -quanta) and particles of a substance (electrons and positrons) takes place.*

The properties of particles differ in many respects from the properties of electromagnetic field (light). The most drastic difference is that all bodies around us are built of particles. It may seem that light only exercises the energy transfer from some bodies to others. For this reason, even at the beginning of the 20th century it was believed that light (electromagnetic field) and matter are separated by an unsurmountable barrier. Later, quantum properties of light were discovered. It turned out that light combines the properties of the flow of particles (photons) with the wave properties. On the other hand, wave properties that had been earlier ascribed only to light as its distinctive feature were discovered in particles of a substance (see Secs. 22.16 and 22.17). These discoveries bridged the gap in our concepts of light and matter. The more so, after the discovery of mutual conversions of light ( $\gamma$ -quanta) and particles of a substance (electron-positron pairs) it has become clear that there is a deep-rooted unity between light and matter: particles of substance and photons (electromagnetic fields) are two different forms of matter. As was mentioned earlier (see Sec. 22.7), photons exhibit many features in common with other particles but have an important peculiarity in that *their rest mass is equal to zero*. A photon always moves at a velocity of light. When it comes to a halt (like during absorption) light does not exist any longer.

#### 24.7. Application of Einstein's Law to Annihilation and Pair Formation

According to (22.6.1), a particle at rest has an internal (rest) energy equal to  $mc^2$ , where  $m$  is the rest mass of the particle.

When an electron and a positron at rest annihilate, their rest energies are completely converted into the electromagnetic energy of two  $\gamma$ -quanta. The rest energies of the electron and the positron are equal to  $mc^2$  each, where  $m = 0.911 \times 10^{-30}$  kg. The energy of each  $\gamma$ -quanta is  $h\nu$ . According to the energy conservation law, we have

$$\begin{aligned} \text{i.e.} \quad & 2h\nu = 2mc^2, \\ h\nu = mc^2 &= 0.911 \times 10^{-30} \times (3 \times 10^8)^2 = 8.2 \times 10^{-14} \text{ J} \\ &= 8.2 \times 10^{-14} / (1.6 \times 10^{-19}) \text{ eV} = 0.51 \times 10^6 \text{ eV} = 0.51 \text{ MeV}. \end{aligned}$$

Thus, the energy of each  $\gamma$ -quantum emitted as a result of the annihilation of an electron and a positron must be equal to 0.51 MeV. The measurement of the energies of the formed  $\gamma$ -quanta gives the results which are in brilliant agreement with this conclusion.

When an electron-positron pair is formed from a  $\gamma$ -quantum, the energy  $h\nu$  of the  $\gamma$ -quantum is converted into the rest energy and the kinetic energy of the particles. Applying the energy conservation law, we obtain

$$h\nu = 2mc^2 + W_k,$$

where  $W_k$  is the total kinetic energy of the electron-positron pair.

Using previous calculations, we can write

$$h\nu = 2 \times 0.51 + W_k = 1.02 \text{ MeV} + W_k.$$

Since kinetic energy is always positive, the pair may be formed only from  $\gamma$ -quanta having an energy higher than 1.02 MeV. Experiments confirm this result as well as the relation between the energy of a  $\gamma$ -quantum and the kinetic energy of the electron-positron pair, obtained earlier.

Thus the analysis of annihilation and pair formation confirms the validity of Einstein's law.

#### 24.8. The Structure of Atomic Nuclei

It was mentioned above (see Sec. 22.8) that the masses of atoms, and hence the masses of atomic nuclei, are always close to an integral number of atomic mass units. This suggests that atomic nuclei are built of particles having a nearly unit mass. Such particles are the proton and the neutron.

At first glance, it seems that besides neutrons and protons, nuclei must also contain positrons and electrons since many nuclei (those of radioactive isotopes) emit these particles. However, a detailed analysis of various properties of nuclei indicates that positrons and electrons as such are absent in nuclei.

For example, some artificial radioactive substances (like the copper isotope  $^{64}\text{Cu}$ ) emit two species of particles, viz. positrons and electrons. A fraction of atomic nuclei of such a substance is converted as a result of decay into the previous element of the Periodic Table, emitting a positron, while the other fraction of nuclei of the same substance is converted into the next element, emitting an electron. It would seem that nuclei of such a substance must contain both positrons and electrons.<sup>6</sup> But the simultaneous existence of electrons and positrons within the volume of a nucleus contradicts the property of these particles to combine and give birth to a pair of  $\gamma$ -quanta.

---

<sup>6</sup> Experiments reveal a complete coincidence of all properties (charge, mass, etc.) of all atomic nuclei of a given isotope. Such a complete coincidence points to the identity of such nuclei. Therefore, we cannot assume that some of  $^{64}\text{Cu}$  nuclei contain electrons while others contain positrons.



But if positrons and electrons are not present as such in a nucleus, in the process of decomposition of the nucleus, which is accompanied by the emission of one of these particles, the latter are formed anew as a result of energy transformations within the nucleus. If a positron (positive charge) is emitted in this case, one of protons is transformed into a neutron, while the emission of an electron (negative charge) indicates that one of neutrons has become a proton. The assumption concerning the formation of electrons and positrons during radioactive decay is hence quite natural since, as was mentioned in Sec. 24.6, these particles are formed in other processes as well.

The hypothesis about the proton-neutron composition of atomic nuclei was put forth by the Soviet physicist D. D. Ivanenko (b. 1904) and the German physicist Werner Heisenberg (1901-1976) soon after the neutron had been discovered. The validity of the proton-neutron model of nucleus was proved by the works of many scientists.

Since the mass number of the proton and of the neutron is unity, the mass number of a nucleus is equal to the total number of particles (protons and neutrons) constituting it. The nuclear charge expressed in elementary units of charge is obviously equal to the number of protons in the nucleus. Thus, according to the proton-neutron model, an atomic nucleus having a mass number  $A$  and an atomic number  $Z$  contains  $A$  particles among which there are  $Z$  protons and  $A - Z$  neutrons. For example, the oxygen nucleus  $^{16}_8\text{O}$  consists of 8 protons and  $16 - 8 = 8$  neutrons. The nucleus  $^{206}_{82}\text{Pb}$  of the lead isotope contains 82 protons and  $206 - 82 = 124$  neutrons, and so on.

The simplest atomic nucleus, viz. a proton, is that of hydrogen. By adding a neutron to the proton, we obtain the simplest compound nucleus, viz. deuteron, or the nucleus of heavy hydrogen (denoted by  $^2_1\text{H}$ , or  $\text{D}$ ).

By adding another neutron, we form the nucleus of a still heavier hydrogen isotope known as *tritium* ( $^3_1\text{H}$ , or  $\text{T}$ ). Tritium is an artificial radioactive substance. It decays with a half-life of 12 years, emitting electrons. As a result of tritium decay, a nucleus with a mass number 3 and a charge 2 is formed. It is the light isotope of helium  $^3_2\text{He}$  consisting of two protons and a neutron. This isotope is stable and is contained in very small amounts in natural helium. The nucleus of the main helium isotope  $^4_2\text{He}$  ( $\alpha$ -particle) is formed by adding one more neutron. Therefore, an  $\alpha$ -particle contains two protons and two neutrons.

By increasing further the number of neutrons and protons in nuclei, we obtain all atomic nuclei existing in nature. The composition of light nuclei (up to oxygen) is represented in Fig. 401. It can be seen from this diagram that stable (nonradioactive) light nuclei contain approximately equal numbers of neutrons and protons.

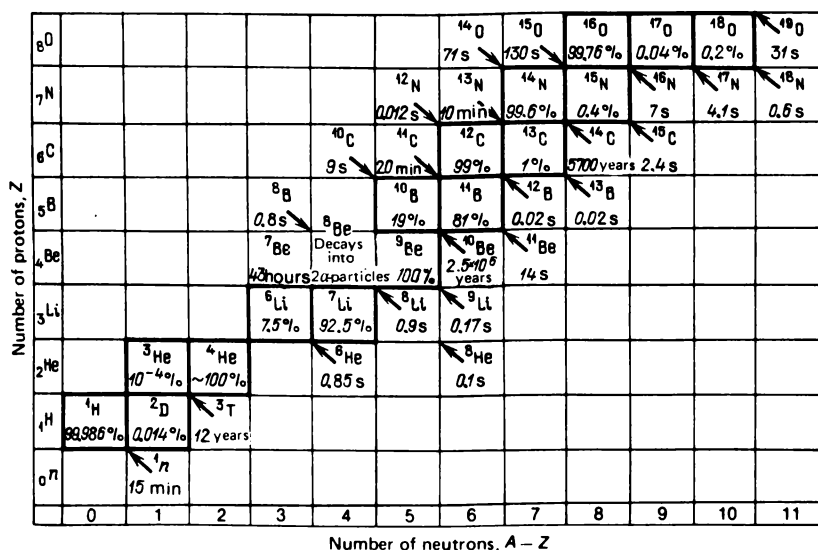


Fig. 401.

Diagram of atomic nuclei (up to oxygen). Solid lines embrace stable isotopes. Cells contain the percent content for stable isotopes and half-lives for radioactive isotopes. Arrows indicate the nucleus into which a radioactive isotope is transformed.

In heavy nuclei, the equilibrium is shifted towards the number of neutrons. For example, in the lead nucleus the number of neutrons is about 1.5 as large as the number of protons. The neutron-to-proton ratio in stable nuclei is the most favourable, imparting the maximum strength to a nucleus. A deviation from such a ratio makes the nucleus unstable. If a nucleus contains too many neutrons, one of them is converted into a proton, i.e. the nucleus decays, emitting an electron (for example,  $^{10}_4\text{Be} \rightarrow ^{10}_5\text{B} + e^- + \nu$ ). On the contrary, if there is an excess of protons, one of the protons is converted into a neutron, emitting a positron (for example,  $^{10}_6\text{C} \rightarrow ^{10}_5\text{B} + e^+ + \nu$ ).

Since nuclear particles (protons and neutrons) are strongly bound in nuclei, attractive forces must act between them. These forces must be strong enough to counteract huge forces of mutual electrostatic repulsion of protons brought together to distances of the order of atomic nuclei ( $10^{-14}$ - $10^{-15}$  m). *Peculiar forces emerging when particles (protons and neutrons) are brought together to small distances and binding these particles in nuclei were called nuclear forces.*<sup>7</sup>

<sup>7</sup> Investigations into nuclear structure show that the attractive nuclear forces act between any pair of particles (two protons, two neutrons, a proton and a neutron). For detail, see Sec. 25.3.

We spoke of nuclear forces while analyzing the capture of slow neutrons by nuclei (Sec. 24.4). Nuclear forces are manifested also in other nuclear reactions and in radioactivity. Although many properties of atomic nuclei have been studied in detail by now, exact laws describing their action remain unclear so far. The establishment of these laws is one of the main problems of modern atomic physics.

### 24.9. Nuclear Energy. Energy Sources of Stars

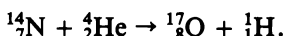
In this course of physics, we came across various forms of energy that can be converted into one another. They include the kinetic energy of moving bodies, the potential energy of bodies in gravitational fields, the energy of electromagnetic fields, the internal energy of bodies and so on. An analysis of nuclear transformations reveals the existence of one more form of energy, viz. *nuclear energy*.<sup>8</sup> *Nuclear energy is the energy stored in atomic nuclei and transformed into other kinds of energy during nuclear transformations, viz. radioactive decay of nuclei and nuclear reactions.* Nuclear energy is manifested in all nuclear transformations.

Let us consider by way of an example nuclear reactions of  $\alpha$ -particles with beryllium and nitrogen described in Secs. 24.1 and 24.2. As a result of reaction (24.3.1), a carbon nucleus and a neutron are formed:



Measurements show that the kinetic energy of reaction products is higher (by 5.7 MeV) than the kinetic energy of the initial nuclei. Consequently, the transformation of the latent nuclear energy into the kinetic energy occurs in this reaction.

In reaction (24.1.1), we have



It turns out that the total kinetic energy of the oxygen nucleus ( ${}^{17}_8\text{O}$ ) and the proton ( ${}^1_1\text{H}$ ) is lower (by 1.2 MeV) than the kinetic energy of the  $\alpha$ -particle causing the reaction (the nitrogen nucleus was at rest at the initial moment). Thus, in this reaction, on the contrary, the kinetic energy is transformed into the nuclear energy. The store of the latter type of energy in reaction products is larger than in the initial nuclei.

The nuclear energy converted into the kinetic energy and back can be calculated if we know the exact values of the masses of the nuclei participating in a reaction. Indeed, according to the energy conservation law, the increment of the kinetic energy is equal to a decrease in the internal energy of the nuclei. But according to Einstein's law, the decrease in the internal energy is equal to the difference in the rest masses of the initial and final reaction products,

---

<sup>8</sup> Instead of the term "nuclear energy", a less accurate term "atomic energy" is often used.

multiplied by  $c^2$ . Let us consider, for example, reaction (24.9.1). The masses of particles participating in the reaction are given below.

| Particle            | Mass <sup>9</sup> , amu | Total mass, amu |
|---------------------|-------------------------|-----------------|
| ${}^9_4\text{Be}$   | 9.0150                  | 13.0189         |
| ${}^4_2\text{He}$   | 4.0039                  |                 |
| ${}^{12}_6\text{C}$ | 12.0038                 | 13.0128         |
| $n$                 | 1.0090                  |                 |

The mass of the initial particles is larger than the mass of the end products by  $13.0189 - 13.0128 = 0.0061$  amu. The internal energy of the particles decreases as a result of the reaction by<sup>10</sup>

$$0.0061 \times 930 = 5.7 \text{ MeV.}$$

As was mentioned above, direct measurements show that the kinetic energy of the reaction products exceeds the kinetic energy of the initial nuclei by just this value (5.7 MeV). Thus, we have another proof of the validity of the relation  $E_0 = mc^2$ .

The nuclear energy is similar to the chemical energy in that the two kinds of energy are manifested in transformations of particles. The chemical energy is manifested in transformations of molecules (i.e. in chemical reactions), while the nuclear energy, in transformations of atomic nuclei (i.e. nuclear reactions). The nuclear and chemical energies have a drastic difference in scale. The energy of chemical reactions is measured in electronvolts (for example, the burning of carbon is accompanied by the liberation of 4.1 MeV of energy per  $\text{CO}_2$  molecule). The energy of nuclear transformations is measured in megaelectronvolts, i.e. is million times higher in the order of magnitude. The larger scale of nuclear processes is due to the huge values of nuclear forces (see Sec. 24.8).

Nuclear transformations in which the latent nuclear energy is transformed into other kinds of energy play an important role in nature. Since 1940's, it has become significant in engineering also. The simplest of such transformations are phenomena involving radioactive decay. As was noted in Sec. 23.5, the energy of radioactive radiation is ultimately converted into heat. Radioactive heat has played the major role in geological processes

<sup>9</sup> The first three rows of the table contain the rest masses of neutral atoms. Their values have been obtained from mass-spectroscopic measurements. The numbers of electrons on the right- and left-hand sides of the equation of a nuclear reaction are equal. Therefore, while calculating the difference in masses, the electron masses are cancelled out, and we obtain the differences in the masses of the nuclei.

<sup>10</sup> The rest energy per atomic mass unit is 930 MeV:

$$\begin{aligned} 1 \text{ amu} \cdot c^2 &= 1.66 \times 10^{-27} (3 \times 10^8)^2 = 1.49 \times 10^{-10} \text{ J} \\ &= 1.49 \times 10^{-10} / (1.6 \times 10^{-19}) \text{ eV} = 9.3 \times 10^8 \text{ eV} = 930 \text{ MeV.} \end{aligned}$$

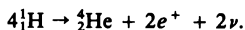
since the decay of uranium, thorium and potassium contained in the Earth's crust is the source of energy which maintains the high temperature in the interior of the Earth. However, the role of natural radioactivity as a commercial source of power is negligibly small: all available radioactive elements decay too slow, and the ways to accelerate their decay are unknown so far.

Unlike in radioactivity, the energy liberation rate in nuclear reactions may vary over a wide range, and the liberated energy may attain huge values. Nuclear reactions are the only available source whose energy is sufficient to sustain radiation of stars during their lifetime, i.e. over billion years. Astrophysical data indicate that the temperature in the interior of stars is of the order of millions or tens of million degrees. At such temperatures, atoms are ionized almost completely. The stellar matter is in the state known as plasma, i.e. has the form of a gas consisting of electrons and "bare" atomic nuclei moving at random at huge velocities. The velocities of random motion are so high that in spite of the electrostatic repulsion of charged nuclei, they experience collisions resulting in nuclear reactions.

At a sufficiently high initial temperature of a star, the number of particles participating in reactions will be very large. The inflow of the liberated nuclear energy compensates energy losses on luminous radiation, and the star will be heated instead of being cooled. In this case, an initiated nuclear reaction ensures the conditions for its continuation (i.e. maintains the high temperature of the medium). Therefore, it will last until the source of "nuclear fuel" is exhausted, i.e. until all nuclei capable of reacting are used up.

"Nuclear fuel" can be beryllium in combination with helium (reaction (24.9.1)), lithium, heavy hydrogen and other substances. But all these substances are contained in stars in relatively small amounts and may serve as individual energy sources only over comparatively short stages of star evolution. It is assumed at present that the main type of "nuclear fuel" that can ensure the energy stock of a star for many billion years is hydrogen.

Hydrogen is the main component of stellar matter. Experiments and the theory of nuclear reactions indicate that hydrogen must be transformed into helium as a result of a series of consecutive nuclear reactions. The total result of these reactions is expressed by the following reaction equation:



In other words, four protons form a helium nucleus, two positrons and two neutrinos. The liberated energy amounts to (taking into account the annihilation of positrons) about 27 MeV, i.e. about 650 billion joules per gram of hydrogen (!).

According to modern views the conversion of hydrogen into helium is the source of energy of stars, including the Sun. It can easily be calculated that hydrogen expenditures of the Sun over 100 years amount to only one billionth of the mass of the Sun<sup>11</sup>.

---

<sup>11</sup> The power emitted by the Sun is assumed to be  $3.8 \times 10^{26}$  J/s, and the mass of the Sun is  $2.0 \times 10^{30}$  kg.

A nuclear reaction proceeding at the expense of the high temperature of the medium is known as a *thermonuclear* reaction. A question arises as to how thermonuclear reactions are initiated in stars. A possible cause of the initial heating that “triggers” a reaction is the compression of stellar matter under the action of gravitational forces, i.e. the conversion of the potential energy of gravitation into the internal energy.

The liberation of large amounts of energy under terrestrial conditions remained for a long time an impossible dream. No methods of attaining huge temperatures (million degrees) required for initiating a thermonuclear reaction were known at that time. The utilization of particles accelerated by accelerators did not seem promising. As was mentioned in Sec. 24.2, fast charged particles moving in a medium spend their to ionize and excite atoms and cause nuclear reactions only with a low probability. Therefore, the energy spent for the initial acceleration of particles exceeds the energy gained from a nuclear reaction.

The situation radically changed in 1939 when the investigation into the properties of neutrons was crowned by the discovery of a new nuclear reaction, i.e. a nuclear *fission* reaction established by the German chemists Otto Hahn (1879-1968) and Fritz Strassmann (1902-1980).

#### 24.10. Uranium Fission. Chain Nuclear Reaction

In the experiments on the bombardment of uranium by neutrons it was found that a uranium nucleus splits under the action of neutrons into two nuclei (fragments) having approximately equal masses and charges. This process is accompanied by the emission of several (two or three) neutrons (Fig. 402). Besides uranium, some other elements at the end of the Periodic Table undergo fission. Like uranium, these elements split not only under the action of neutrons, but also without any external influence (spontaneously)<sup>12</sup>. Spontaneous fission was discovered experimentally in 1940 by the

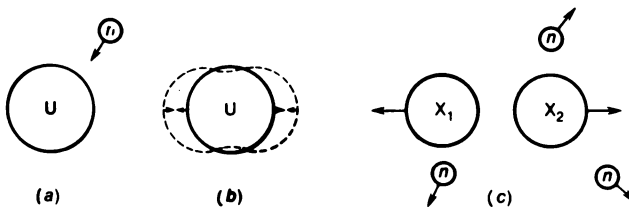
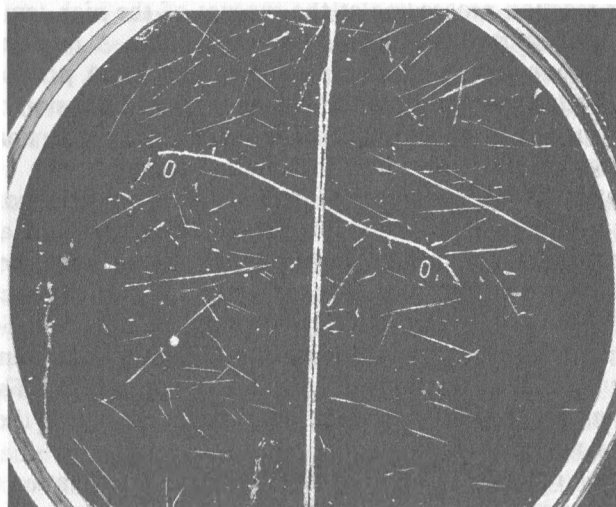


Fig. 402.

Fission of a uranium nucleus under the action of neutrons: (a) the nucleus captures a neutron, (b) the collision of the neutron with the nucleus causes its vibrations, (c) the nucleus is split into two fragments with the emission of a few more neutrons.

<sup>12</sup> Spontaneous fission should be included into the class of radioactive decay processes described in the previous chapter. Like other types of decay, spontaneous fission is an intranuclear process occurring without any external influence.



**Fig. 403.**

Photograph of tracks of uranium fission fragments in a Wilson cloud chamber: fragments (O) fly apart in opposite directions from a thin uranium layer deposited on the partition of the chamber. The photograph shows a large number of thinner tracks belonging to the protons knocked out by neutrons from water molecules of the vapour filling the chamber.

Soviet physicists G. M. Flerov and K. A. Petrzhak. It is quite a rare process. For example, only about 20 spontaneous acts of fission per hour occur in 1 g of uranium.

Due to mutual electrostatic repulsion, the fission products (fragments) fly apart in opposite directions, having acquired a huge kinetic energy of about 160 meV. Therefore, a fission reaction occurs with a considerable liberation of energy. Fast fragments intensely ionize the atoms of a medium. This property of fragments is used for detecting fission processes with the help of an ionization chamber of a Wilson cloud chamber. A photograph of fragment tracks in the Wilson chamber is shown in Fig. 403. Of great importance is the fact that the neutrons emitted as a result of uranium nuclear fission (the so-called *secondary fission neutrons*) are able to cause fission of new uranium nuclei. Therefore, a *chain fission reaction* can be realized: having been initiated once, the reaction may in principle continue by itself, embracing larger and larger number of nuclei. The diagram illustrating the evolution of such a reaction is shown in Fig. 404.

It is difficult to realize a chain fission reaction in practice. Experiments show that no chain reaction emerges in the bulk of natural uranium. The reason behind this is that the secondary neutrons are missing from the reaction. In natural uranium, most of neutrons come out of play without caus-

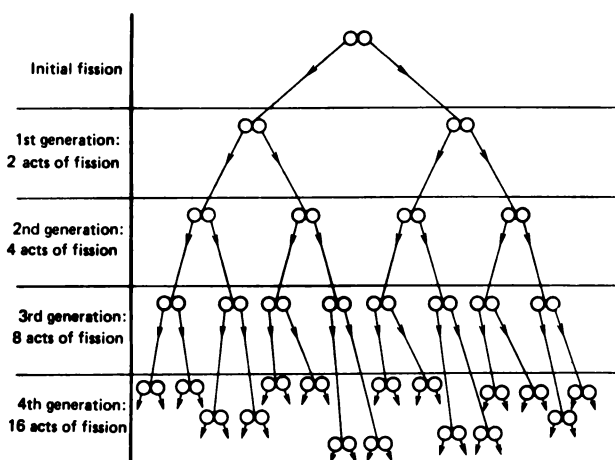


Fig. 404.

Evolution of a chain fission reaction: it is assumed conditionally that nuclei emit two neutrons during fission and there is no loss of neutrons, i.e. each neutron causes a new act of fission. Circles indicate fission fragments and arrows, fission neutrons.

ing fission. It was found out that neutrons are ions in the most abundant uranium isotope, viz. uranium-238 ( $^{238}_{92}\text{U}$ ). This isotope readily absorbs neutrons in a reaction similar to the reaction of silver with neutrons (see Sec. 24.5). As a result, an artificial radioactive isotope  $^{239}_{92}\text{U}$  is formed. The fission of  $^{238}\text{U}$  occurs with difficulty and only under the action of fast neutrons.

The isotope  $^{235}\text{U}$  has the most favourable properties for a chain reaction. The amount of this isotope in natural uranium is 0.7%. Its fission occurs under the action of neutrons of any energy (fast or slow), the rate of fission being the higher, the lower the energy of neutrons. Unlike in  $^{238}\text{U}$ , a simple absorption of neutrons competing with fission, is hardly probable in  $^{235}\text{U}$ . Therefore, a chain fission reaction in pure uranium-235 is possible, however, only if the mass of  $^{235}\text{U}$  is large enough. In a small mass of uranium a fission reaction terminates due to the escape of secondary neutrons beyond the limits of the substance.

Indeed, in view of the very small size of atomic nuclei, a neutron traverses a considerable distance (of the order of centimeter) in a substance before it accidentally comes across a nucleus. If the size of a body is small, the collision probability within the body is very low. Almost all the secondary neutrons escape through the surface of the body without causing new acts of fission, i.e. without sustaining the reaction.

The neutrons formed in the surface layer mainly escape from a body of a large size. Neutrons formed in the bulk of the body have a considerably



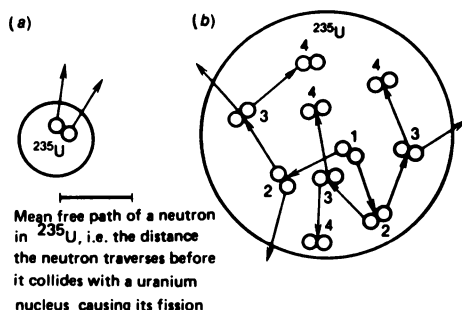


Fig. 405.

Evolution of the chain fission reaction in  $^{235}\text{U}$ . (a) In a small mass of  $^{235}\text{U}$ , most of fission neutrons escape from the reactor. (b) In a large mass of uranium, many fission neutrons cause the fission of new nuclei, and the number of fission acts grows from generation to generation. Circles indicate fission fragments and arrows, fission neutrons.

thick uranium layer in front of them, and most of them cause new acts of fission, thus sustaining the reaction (Fig. 405). The larger the mass of uranium, the smaller fraction of the volume is occupied by the surface layer<sup>13</sup> that loses many neutrons, and the more favourable are the conditions for the development of a chain reaction.

By gradually increasing the amount of  $^{235}\text{U}$ , we attain the *critical mass*, i.e. the minimum mass starting from which a nondecaying chain reaction of fission is possible in  $^{235}\text{U}$ . With a further increase in the mass of  $^{235}\text{U}$ , the reaction intensity strongly increases (the reaction can be initiated by spontaneous fission). If the mass of  $^{235}\text{U}$  is decreased to a value smaller than the critical, the reaction terminates.

Therefore, a chain fission reaction can be realized if we have a sufficiently large amount of pure  $^{235}\text{U}$  separated from  $^{238}\text{U}$ .

As was shown in Sec. 22.9, the isotope separation is a complicated and expensive but still feasible problem. Indeed, the extraction of  $^{235}\text{U}$  from natural uranium is a method used for the practical realization of the chain fission reaction.

A chain reaction was also realized by another method that does not involve the separation of uranium isotopes. This method is in principle more complicated but easier to realize. It is based on deceleration of fast secondary fission neutrons down to the velocities of thermal motion. It was shown that nondecelerated secondary neutrons in natural uranium are mainly absorbed by the  $^{238}\text{U}$  isotope. Since the absorption in  $^{238}\text{U}$  does not lead to fission, the reaction terminates. Measurements show that when

<sup>13</sup> The area of a spherical surface is proportional to the squared radius, while the volume is proportional to the cube of the radius. Hence the volume of a sphere increases with its radius faster than its surface area.

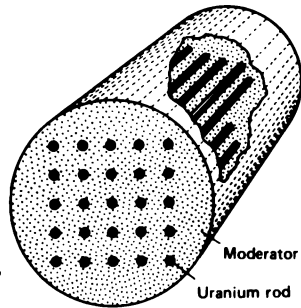


Fig. 406.

A system consisting of natural uranium and a moderator, in which a chain fission reaction may develop.

neutrons are decelerated to thermal velocities, the absorptivity of  $^{235}\text{U}$  increases more than the absorptivity of  $^{238}\text{U}$ . Therefore, the absorption of neutrons by the  $^{235}\text{U}$  isotope, resulting in fission, prevails. This means that *if fission neutrons are decelerated before they are absorbed in  $^{238}\text{U}$ , the chain reaction becomes possible in natural uranium as well.*

In practice, this is realized by placing thin rods of natural uranium in the form of a sparse lattice into a moderator (Fig. 406). Moderators are substances having a small atomic mass and weakly absorbing neutrons. Good moderators are graphite, heavy water and beryllium.

Let us suppose that a uranium nuclear fission has occurred in a rod. Since the rod is relatively thin, almost all fast secondary neutrons escape to the moderator. Since the rods are sparsely arranged in the lattice, an escaping neutron undergoes many collisions with moderator nuclei before it reaches another rod, and is decelerated to the velocity of thermal motion (Fig. 407). After entering another uranium rod, the neutron in all probability will be absorbed by  $^{235}\text{U}$  and will cause a new fission, thus sustaining the reaction. The chain fission reaction was realized for the first time in 1942 in the USA by the group of scientists headed by the Italian physicist

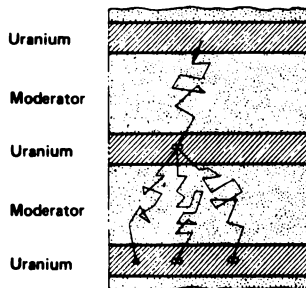


Fig. 407.

Evolution of a chain fission reaction in a system consisting of natural uranium and a moderator. A fast neutron escaping from a thin rod gets into the moderator and is decelerated. Getting into uranium again, the slow neutron in all probability is absorbed by  $^{235}\text{U}$ , causing fission (denoted by two light circles). Some neutrons are absorbed by  $^{238}\text{U}$  without causing fission (dark circle).

E. Fermi (1901-1954) in a system containing natural uranium. Independently, this process was carried out in the USSR in 1946 by the Academician I. V. Kurchatov (1903-1960) with coworkers.

### 24.11. Application of Nondecaying Chain Fission Reaction.

#### Atom and Hydrogen Bombs

The chain fission reaction makes it possible to split considerable amounts of uranium nuclei. This process is accompanied by the liberation of a large amount of energy. Depending on the conditions, a chain reaction can be either a calm, controlled process, or an explosion.

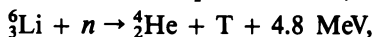
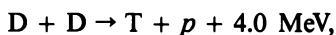
If the mass of a reacting mixture exceeds the critical mass only slightly, the reaction develops slowly. When the required intensity has been attained, the reaction growth can be terminated. For this purpose, it is sufficient to reduce the mass of the system to the critical value. At any moment, the reaction can be extinguished by reducing the mass below the critical value<sup>14</sup>. Thus, the chain reaction can be controlled completely.

Another situation takes place when the mass of a system is considerably larger than the critical mass. Then the reaction develops at the rate of an explosion. Once initiated, it comes out of control, and the rapid evolution of energy leads to the destruction of the system.

The reaction rate is especially high in pure  $^{235}\text{U}$  since it is caused here by fast (nondecelerated) neutrons. For this reason,  $^{235}\text{U}$  in an amount considerably larger than the critical mass is the most powerful explosive used for the so-called *atom bomb*. In order to prevent an atom bomb from exploding during storage, its uranium charge can be divided into several separate portions having a mass smaller than the critical value. For obtaining an explosion, these parts must be rapidly brought together.

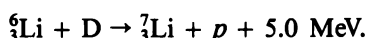
The energy liberated during an explosion of the uranium charge is  $10^5$  times higher than that of ordinary explosives taken in the same amount.

At the moment of explosion, the temperature in an atom bomb rises to million degrees. Therefore, if the explosion of the bomb takes place in an appropriate medium, it may cause a thermonuclear reaction (see Sec. 24.9). Among the substances having the most favourable properties for the evolution of a thermonuclear reaction are heavy hydrogen (deuterium  $^2\text{D}$ ), superheavy hydrogen (tritium  $^3\text{T}$ ), and lithium. For example, the following nuclear reactions occur in a mixture of these substances:




---

<sup>14</sup> The practically used method of controlling the reaction consists in introducing (or extracting) nonfissionable substances that strongly absorb neutrons.



A system incorporating an atom bomb and a substance in which a high-power thermonuclear reaction is initiated by the explosion is known as a *nuclear fission bomb*, or *hydrogen (H-) bomb*. The explosive power of an H-bomb is hundreds of times higher than that of an atom bomb. As a matter of fact, the amount of “explosive” ( ${}^{235}\text{U}$ ) in the atom bomb is limited: the mass of each its part must be smaller than the critical mass to exclude a premature explosion. On the other hand, there is no such limitations for the amount of “explosive” in the H-bomb since neither deuterium, nor tritium (or their mixture) can explode spontaneously.

Unlike the fission reaction, thermonuclear reactions have not been used so far for obtaining thermal and electric energy. But intense investigations in this direction are carried out in the USSR and other countries. The application of nuclear fusion for obtaining energy is of vital importance since the raw material resource for this reaction is unlimited (deuterium is contained in sea water), while uranium deposits are not very large.

In order to initiate a thermonuclear fusion reaction a nuclear “fuel” must be heated to a temperature of the order of ten million degrees. At such temperatures, substances go over to the state of a strongly ionized gas, viz. plasma. In order to sustain a reaction the plasma must be confined to prevent its expansion, i.e. the free motion of plasma particles (ions and electrons) must be bounded. This cannot be attained just by placing the plasma into a closed vessel since no wall can withstand a temperature exceeding thousand times the evaporation point of refractory materials (the other reason necessitating the insulation of plasma from the vessel walls is that the intense heat transfer to the walls would hamper the plasma heating).

At the beginning of the fifties, the Soviet physicists A. D. Sakharov and I. E. Tamm, as well as other physicists abroad, proposed that a strong magnetic field be used for plasma confinement. As was mentioned above (see Sec. 22.5), a charged particle whose initial velocity is perpendicular to the magnetic induction of a uniform magnetic field moves in a circle in a plane perpendicular to the direction of the field. If the initial velocity is parallel to the magnetic field, the particle moves freely (by inertia) along a magnetic field line since in this case the Lorentz force is zero. In the general case when the initial velocity is directed arbitrarily, the rectilinear and circular motions are added, and the particle describes a helical trajectory wound on a magnetic field line (Fig. 408a). Such a form of motion is retained in a nonuniform magnetic field if the direction of the magnetic induction of the field changes insignificantly over a distance of the order of the lead of a “screw” (Fig. 408b). The particle turns out to be as if tied to the field line, since it is kept at the same distance from the line, which is equal to the radius of the helix. The helix radius is directly proportional to the particle's velocity<sup>15</sup> and inversely proportional to the magnetic induction  $B$  (see Sec. 22.5). By increasing  $B$ , we can make the radius of the helix as small as desired.

In a real plasma, the motion of particles is affected by their collisions and by the internal magnetic and electric fields of the plasma (they are always present since plasma consists of charged particles). For this reason, the analysis of the effect of an external magnetic field on the motion of plasma particles turns out to be very complicated. The main feature of

<sup>15</sup> To be more precise, to the velocity component perpendicular to the magnetic field.

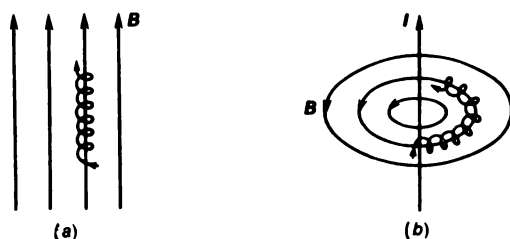


Fig. 408.

Motion of a slow charged particle in a uniform magnetic field (a) and in the magnetic field of a straight current-carrying conductor (b). Arrows indicate the direction of the magnetic field and helical lines, the trajectories of particles.

this effect, however, remains unchanged: bending the trajectories of particles, the magnetic field strongly hampers their motion in the direction perpendicular to the magnetic field lines. This peculiarity is used for the confinement (insulation) of the plasma.

The magnetic field is also employed for plasma heating: an induced emf accelerating ions and electrons emerges as a result of the change in the magnetic induction.

At present, physicists know how to heat the plasma (true, rather rarefied) to hundred million degrees and how to confine it in such a state for hundredths of a second. These achievements give grounds to hopes that ultimately controlled nuclear fusion reaction (and not explosive as in an H-bomb) will be realized in this way.

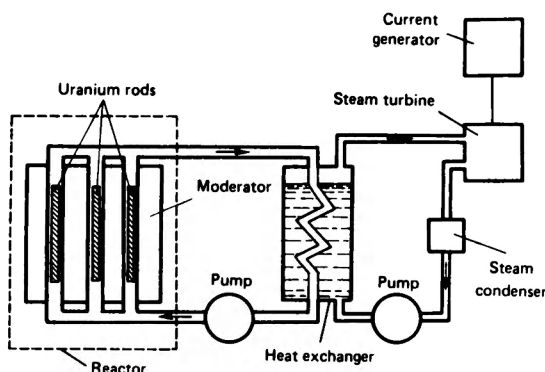
During the explosion of an A- or H-bomb, a large number of neutrons and  $\gamma$ -rays are emitted and large amounts of radioactive substances are formed in addition to the effects typical of any high-power explosion. Radiation from reaction products makes the region of explosion dangerous for life for a long time after the explosion. Radioactive products are carried by air flows over distances of the order of thousand kilometres from the region of explosion. Having detected an elevated radioactivity of air with the help of a counter, the fact of explosion of an A- or H-bomb can be reliably established.

## 24.12. Nuclear Reactors and Their Applications

A device in which a controlled chain fission reaction takes place is called a *nuclear reactor*. Uranium or plutonium (artificially obtained radioactive element with an atomic number  $Z = 94$ ) is used as a substance undergoing fission (nuclear fuel).

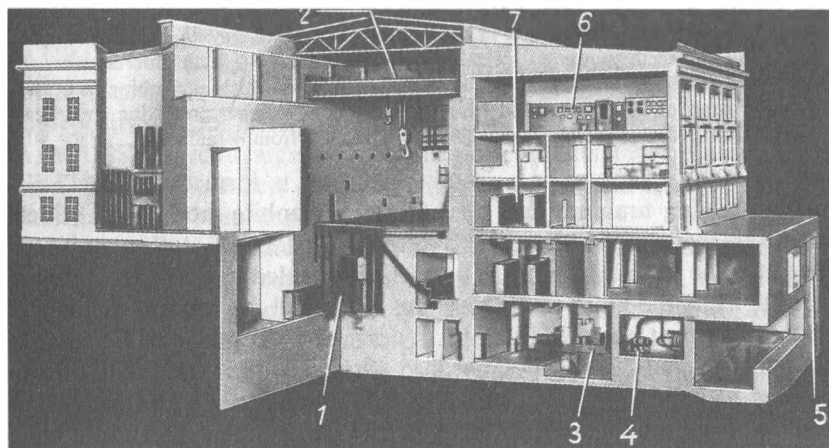
Nuclear reactors are used for producing energy, for obtaining radioactive isotopes (including transuranic elements, i.e. the elements with  $Z > 92$ ) and as sources of intense neutron beams. Let us consider these applications.

**1. Energy production.** Fission fragments are decelerated in uranium over a short distance (less than  $5\mu\text{m}$ ). As a result, almost the entire energy liberated in a reactor evolves in the form of heat in the bulk of uranium. This heat can be used, for example, for heating and vaporizing the liquid flowing

**Fig. 409.**

A lay-out of a nuclear power plant. The uranium rods of the reactor are washed by a heat-transfer agent (gas, water or molten metal) which takes away the heat liberated in the rods and transfers it in the heat exchanger to water converted into steam. As at an ordinary electric power plant, steam drives a steam turbine and the electric generator connected to it. In another version which is also widely used, steam is formed directly in the reactor, and there is no heat exchanger.

past uranium and then converted into the mechanical and electric energy with the help of a steam turbine, or some other heat engine (Fig. 409). The first nuclear power plant based on this principle was constructed in the USSR in 1954 (Fig. 410). The schematic diagram of this power plant is

**Fig. 410.**

General view of a nuclear power plant (1954); 1 — reactor, 2 — crane for the replacement of "burnt" uranium rods, 3, 4 — pump and electric motor ensuring the circulation of water through the reactor, 5 — heat exchanger, 6 — control room (control panel), 7 — panel with instruments indicating inadmissible radioactivity in various departments of the plant.

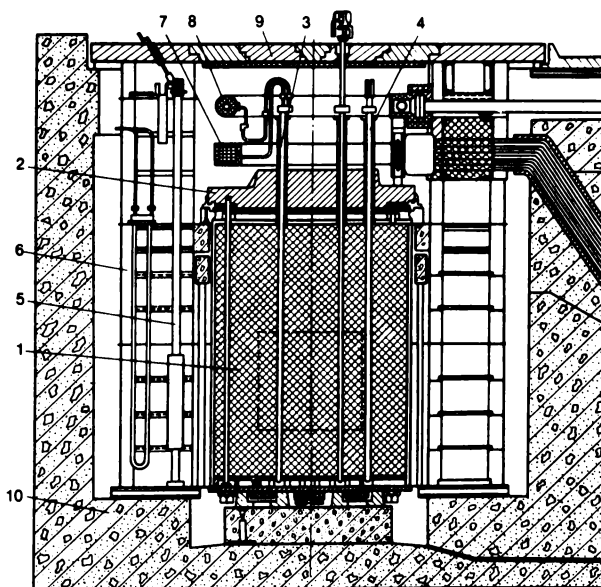


Fig. 411.

Schematic diagram of the reactor of the first Soviet nuclear power plant: 1 — graphite lining of the reactor, enclosed into a hermetic steel shell; broken lines embrace the active zone of the reactor, containing uranium, the remaining graphite being intended for the reflection of neutrons, 2 — upper slab (made of cast iron), 3 — one of 128 working channels in which uranium rods are placed and cooling water flows (under a pressure of 100 atm), 4 — a channel for moving a control rod containing a neutron absorber (boron); control rods are intended for monitoring the reactor power and terminating the reaction, 5 — ionization chamber for measuring the reaction intensity in the reactor, 6 — water shield absorbing neutrons, 7, 8 — water supplied to and removed from the reactor, 9 — upper protecting cover (cast iron), 10 — concrete shield (protecting mainly from  $\gamma$ -radiation).

shown in Fig. 411. The main part of the reactor is formed by “fuel” elements containing uranium and placed into a graphite moderator. “Fuel” elements are made in the form of two thin-walled stainless-steel tubes inserted one into the other. Uranium is hermetically sealed in the space between the tubes, while the inner cavity serves as a channel for running water that takes away the heat liberated in uranium during the operation of the reactor. The hermetic sealing is required because of the chemical instability of uranium and to prevent the leakage of harmful radioactive gases formed as a result of reaction. In order to facilitate the development of a chain reaction, the “fuel” elements are made of uranium artificially enriched by the readily fissionable  $^{235}\text{U}$  isotope (there are 5%  $^{235}\text{U}$  in enriched uranium used in reactors, while natural uranium contains just 0.7% of this isotope).

The operation of a uranium reactor is accompanied by intense radioactivity. In order to protect the personnel from radioactive radiation and neu-

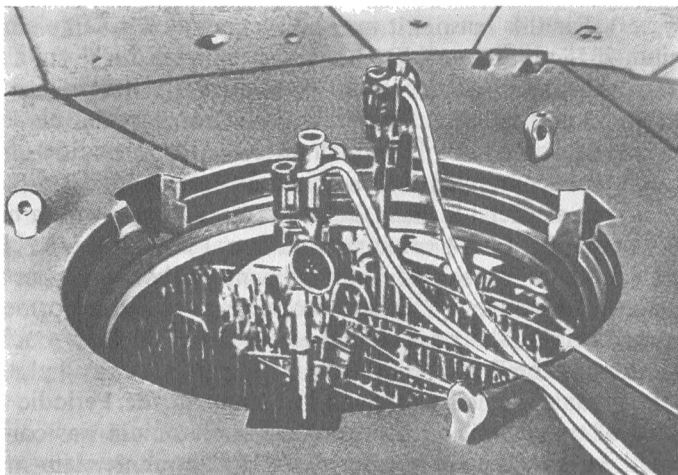


Fig. 412.

The upper part of a reactor without a cover. The motors for driving control rods can be seen on the photograph. Below them are the pipes for supplying water to working channels.

trons which are also very harmful in large doses, the reactor is enclosed by a thick-walled shell made of concrete and other materials (Figs. 411 and 412).

The remarkable feature of a nuclear reactor as a power source is the low fuel expenditure. The amount of heat liberated during fission of 1 g of  $^{235}\text{U}$  is the same as that obtained by burning several tons of coal. This makes it possible to use reactors in the regions located far from coal or oil deposits and for transportation (on ships, submarines and aeroplanes). In the USSR, several large-scale atomic power plants have been put in operation. A few ice-breakers equipped with atomic engines and atomic submarines have also been constructed.

Nuclear power engineering will be still more important in future. It has been calculated that with a modern rate of consumption of power, the shortage of coal and oil will be felt even in 50 years. The utilization of uranium is a way out since the energy stored in uranium deposits is 10-20 times higher than the energy stored in organic fuels. The problem of energy resources will be solved completely after controlled fusion reactions have been realized (see Sec. 24.11).

**2. Transuranic elements.** As a result of bombardment of uranium by neutrons, the  $^{238}_{92}\text{U}$  isotope is converted into  $^{239}_{92}\text{U}$ . This isotope is unstable: as a result of the  $\beta$ -decay, it is transformed into an isotope of neptunium-93 ( $^{239}_{93}\text{Np}$ ). This isotope, in turn, undergoes the  $\beta$ -decay and over a short time (its half-life is 2.35 days) is converted into an isotope of plutonium-94, viz.



$^{239}_{94}\text{Pu}$ . Plutonium-239 is also unstable, but it decays very slowly (its half-life is 24 000 years). For this reason, it may be accumulated in large amounts. Like uranium-235, plutonium-239 is a good “nuclear fuel” suitable for nuclear reactors and for atom bombs. Plutonium is obtained in nuclear reactors based on natural uranium with a moderator. Most of neutrons in such reactors are absorbed by  $^{238}\text{U}$ , which results in the formation of plutonium. Plutonium accumulated in uranium can be separated by chemical methods. Another artificial nuclear fuel is the  $^{233}\text{U}$  isotope with a half-life of 162 000 years, which is not contained in natural uranium. Like plutonium,  $^{233}\text{U}$  is formed as a result of bombardment of thorium by neutrons. Thus, low-fissile substances ( $^{238}\text{U}$  and thorium) can be transformed into a valuable nuclear fuel. This possibility is essential since there is much more  $^{238}\text{U}$  and thorium in the Earth’s crust than  $^{235}\text{U}$ . Neptunium and plutonium represent transuranic elements following uranium in the Periodic Table.

The series of transuranic elements following plutonium was continued to an element with an atomic number of 107. Transuranic elements have not been discovered in nature since all of them are radioactive and have a short lifetime in comparison with the geological age of the Earth.

**3. Obtaining radioactive substances.** In an operating reactor, intense neutron flows formed as a result of fission are observed. Various artificial radioactive isotopes can be obtained in a reactor by bombarding substances with neutrons (cf. reaction (24.5.1)). Another source of radioactivity in a reactor are fragments of uranium fission most of which are unstable.

Artificial radioactive elements have found a wide application in science and technology. Substances emitting  $\gamma$ -rays are used instead of a more expensive radium for examining thick metal bodies in transmitted light, for curing cancer, and so on. The property of large doses of  $\gamma$ -radiation to kill living cells of microorganisms are used in preserving foodstuffs. Radioactive radiation is now used in chemical industry since it facilitates many important chemical reactions. The most important is the *method of labelled atoms*. This method is based on that a radioactive isotope is indistinguishable from stable isotopes of the same element as regards its chemical and many physical properties. At the same time, a radioactive isotope can be recognized by its radiation (for example, by using a gas-discharge counter). By adding a radioactive isotope to an element under investigation and detecting its radiation, one can trace the path of this element in an organism in a chemical reaction, during melting of metals, and so on.

**The importance of nuclear power.** The methods of utilization of nuclear energy have been developed quite recently, but the first results obtained by using these methods are significant. Undoubtedly, further development of the methods of production and utilization of nuclear energy will open new unprecedented opportunities for science, engineering and industry. It

is difficult to visualize the full scale of these opportunities at the modern stage. The liberation of nuclear energy indicates an unlimited broadening of man's power provided that the nuclear energy will be employed for peaceful purposes. The Soviet Union has always pursued a policy of using the atomic energy only for peaceful purposes and of prohibition of atomic and nuclear weapons and other weapons of mass destruction.

It should also be noted that the designing of nuclear reactors is one of the most significant results of the theory of the internal structure of matter. The radiation emitted by invisible and untangible atoms and atomic nuclei has lead to quite a visual practical result, i.e. the liberation and utilization of nuclear energy contained in latent form in uranium. This result convincingly proves that our scientific ideas about atoms and atomic nuclei are correct, i.e. reflect the objective reality of nature.

? **IV.36.** Write symbolically the following nuclear reactions: (a) collision between two deuterons as a result of which two particles are formed, the lighter of which being a proton, (b) the same reaction, but the lighter particle is a neutron (symbol  $n$ , the mass is equal to unity and the charge is zero), (c) collision of a proton with the nucleus of a lithium isotope having a mass of 7, leading to the formation of two  $\alpha$ -particles, (d) collision of a deuteron with an aluminium nucleus, leading to the formation of a new nucleus and a proton.

**IV.37.** Why cannot  $\alpha$ -particles emitted by radioactive substances cause nuclear reactions in heavy elements although they initiate such reactions in light elements?

**IV.38.** Nitrogen has been bombarded for an hour by a beam of  $\alpha$ -particles accelerated in a cyclotron. Determine the amount of  $^{17}\text{O}$  formed if the beam current is 200  $\mu\text{A}$  and the nuclear reaction (24.1.1) is initiated by one of 100 000  $\alpha$ -particles in the beam.

**IV.39.** Write the following nuclear reactions: (a) the splitting of a deuteron into a proton and a neutron by a  $\gamma$ -quantum, (b) the capture of a neutron by a proton, accompanied by the emission of a  $\gamma$ -quantum, (c) the splitting of a  $^9\text{Be}$  nucleus by a  $\gamma$ -quantum resulting in the formation of two  $\alpha$ -particles, (d) the capture of a neutron by a nucleus of a nitrogen isotope having a mass number of 14, accompanied by the emission of a proton, and (e) the collision of a beryllium nucleus with a deuteron followed by the emission of a neutron.

**IV.40.** A beam of fast neutrons passes through a 1-cm thick iron plate. Determine the fraction of neutrons undergoing collisions with iron nuclei if the radius of an iron nucleus is  $6 \times 10^{-15} \text{ m}$ .

*Hint.* The required value is equal to the fraction of the surface area of the plate covered by nuclei.

**IV.41.** Using the laws of energy and momentum conservation for an elastic collision between two balls, calculate the fraction of energy lost in a *head-on* collision of a neutron with a stationary nucleus having a mass number of  $A$  amu. Calculate the maximum energy lost by a neutron colliding with a proton, a carbon nucleus, and a lead nucleus.

**IV.42.** Depending on the type of a collision (head-on or glancing), a neutron colliding with a proton loses a certain fraction of its energy. *On the average*, the energy of a neutron colliding with a stationary proton is reduced by half. Determine the *mean* energy of a neutron after  $n$  collisions with protons.

**IV.43.** Determine the *mean* number of collisions with protons required to reduce the energy of a neutron from 1 Mev to 1 eV (see Ex. IV.42).

**IV.44.** Three identical silver plates were bombarded by neutrons under similar conditions but during different time intervals: 1 min, 1 h and 1 day. The measurements of the activity<sup>16</sup> of  $^{108}\text{Ag}$  having a half-life of 2.3 min showed that the activity of the second plate is several times higher than that of the first plate, while the activity of the third plate is the same as the activity of the second plate. Explain the result of experiments.

**IV.45.** In a Wilson cloud chamber partitioned by a solid plate, a track of a particle crossing the plate has been detected (Fig. 413). What is the direction of motion of the particle? What is the sign of its charge if the magnetic field lines are perpendicular to the plane of the figure and directed upwards?

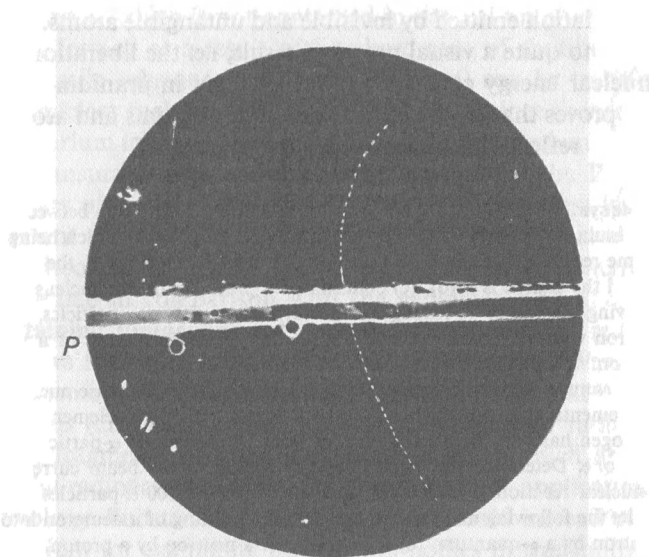
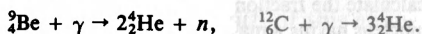


Fig. 413.

To Exercise IV.45. The track of a charged particle in a Wilson cloud chamber. The particle has crossed plate  $P$ . The chamber was placed in a magnetic field whose lines are directed upwards along the normals to the plane of the figure.

**IV.46.** Why do radioactive substances obtained by bombarding stable nuclei by  $\alpha$ -particles undergo the  $\beta$ -decay if the initial reaction is accompanied by the emission of protons and the positron decay if the initial reaction occurs with the emission of neutrons?

**IV.47.** Determine the minimum energy of  $\gamma$ -quanta required for splitting beryllium and carbon nuclei in the reactions



The values of masses of the particles participating in the reactions can be taken from the Table on p. 481.

**IV.48.** A  $^{238}_{92}\text{U}$  nucleus emitting an  $\alpha$ -particle with an energy of 4.2 MeV is transformed into a  $^{234}_{90}\text{Th}$  nucleus. Determine the mass of a  $^{234}_{90}\text{Th}$  atom if the mass of a  $^{238}_{92}\text{U}$  atom is 238.1249 amu. The mass of a  ${}^4_2\text{He}$  atom is given in the table on p. 481.

<sup>16</sup> See footnote on p. 456.

**IV.49.** The highest accuracy with which the mass of an atom or molecule can be measured is one millionth of amu (0.000001 amu). Can Einstein's law be used under this condition for calculating the energy liberated in chemical reactions by measuring the masses of particles participating in a reaction? (The energy liberated in a chemical reaction does not exceed 10 eV.)

**IV.50.** What particles (positrons or electrons) will be emitted by the fragments of fission of  $^{235}\text{U}$  if one of the fragments is  $^{140}\text{Ba}$ ? (Natural barium consists of isotopes with masses ranging from 130 to 138 amu, while natural krypton consists of isotopes with masses ranging from 78 to 86 amu.)

**IV.51.** Determine the power of a reactor in which the fission of 1 g of  $^{235}\text{U}$  lasts for a day. The total energy liberated during the fission of a  $^{235}\text{U}$  nucleus should be taken as 185 MeV.

**IV.52.** The kinetic energy of fission fragments is 160 MeV, the energy of fission neutrons is 5 MeV, and the energy of  $\beta$ -radiation, 10 MeV.

What is the approximate fraction of energy, liberated in a reactor consisting of a moderator and thin uranium rods, evolved in uranium and in the moderator?

**IV.53.** In which of the two cases is the critical mass of uranium in a reactor smaller: when the reactor is surrounded by air, or when it is in a dense substance that weakly absorbs neutrons?

**IV.54.** A fraction of secondary neutrons emitted during uranium fission in a reactor perishes without causing new acts of fission (escapes from the reactor or is captured by nuclei of the reactor material), while the other fraction causes new fissions of uranium nuclei. The number of acts of fission caused by secondary neutrons emitted during the fission of a uranium nucleus is known as the *multiplication factor*  $k$  of the reactor. It shows how many times the number of fissions increases during the lifetime of *one generation* of neutrons.

How does the reactor power vary with time if (a)  $k > 1$ , (b)  $k = 1$ , and (c)  $k < 1$ ?

**IV.55.** About 2.5 secondary neutrons are emitted on the average during fission of a uranium nucleus. In the reactor of a nuclear power plant, 50% of all fission neutrons are captured without causing fission. Determine the fraction of fission neutrons escaping from the reactor if the latter operates at a constant power.

**IV.56.** The mean lifetime of one generation of neutrons in a reactor is about 0.1 s, the multiplication factor being 1.005. Determine the time  $t$  during which the reactor power increases from  $10^{-6}$  to  $10^6$  W.

**IV.57.** Petrol and petroleum are successively pumped through the same pipeline. Suggest a method for determining the instant of time at which the petrol-petroleum interface passes through a given cross section of the pipeline, without taking a test sample.

## Chapter 25

# Elementary Particles

### 25.1. General Remarks

Analyzing the structure of matter, we come to the conclusion that the whole variety of bodies surrounding us is essentially constituted by three types of very small particles, viz. protons, neutrons and electrons. Protons and neutrons combined in groups form atomic nuclei. Together with electrons, nuclei form atoms, atoms are combined into molecules, and so on.

At the modern stage of development of science, it is unclear whether electrons, protons and neutrons are the simplest undecomposable particles or, like atoms they are constructed of other (yet unknown) simpler particles. The particles about which we cannot say for certain that they are compound are known as *elementary particles*. Besides electron, proton and neutron, other elementary particles are also known.

Among them, we must mention first of all electromagnetic quanta (photons). Photons do not constitute atoms, molecules, etc. directly, but they play an important role in nature as agents transporting energy from some particles to others or from some bodies to others. The properties of electrons, protons and neutrons differ from the properties of photons. However, as was mentioned in Secs. 22.16, 24.6 and 24.7, there exists a deep similarity between photons and other elementary particles. This allows us to assume that photons form a type of elementary particles.

---

<sup>1</sup> During last two decades that elapsed since this chapter had been written, the elementary particle physics has undergone radical changes, and the information about elementary particles contained in this book is far from being complete. For this reason, we decided to confine ourselves to the minimum refinement of the material presented in this chapter and to supplement it with a short new Chapter 26 containing the latest advances in this important branch of modern physics in concise form.

Chapters 25 and 26, unlike the other chapters of this textbook, employ the units adopted in scientific literature on elementary particles. For example, the unit of length is 1 cm or 1 fermi ( $10^{-13}$  cm), the unit of energy is 1 eV (1 keV =  $10^3$  eV, 1 MeV =  $10^6$  eV, 1 GeV =  $10^9$  eV, 1 TeV =  $10^{12}$  eV). The mass and momentum are expressed in terms of energy units. For example, the mass  $m = 1 \text{ GeV}/c^2$  denotes a mass for which the rest energy is  $mc^2 = 1 \text{ GeV}$ . The electric charge is measured in the units of elementary charge  $e = 1.602 \times 10^{-19}$  C. Planck's constant in these units is  $h = 4.136 \times 10^{-24} \text{ GeV} \cdot \text{c}$ . The following factors are used for transformation to SI units: energy 16 eV =  $1.602 \times 10^{-10}$  J. The momentum 1 GeV/c =  $5.35 \times 10^{-19}$  kg·m/s the mass 1 GeV/c<sup>2</sup> =  $1.783 \times 10^{-27}$  kg = 1.074 amu.

Further, elementary particles include the *positron* and the *neutrino*. The positron is a twin of the electron, differing from it only in the charge sign. We encountered positrons while considering artificial radioactivity and interaction of hard  $\gamma$ -quanta with substances (the formation of electron-positron pairs (see Sec. 24.6). The neutrino is a neutral particle emitted together with electrons or positrons during the  $\beta$ -decay of atomic nuclei. The properties of neutrinos will be considered below.

Finally, mesons and other particles belonging to elementary particles are formed in nuclear reactions caused by particles and high-energy  $\gamma$ -quanta (hundreds of billions electron volts). These particles will be described in Secs. 25.3 and 25.4.

An analysis of the properties of elementary particles shows, first of all, that elementary particles are far from being invariable: they undergo mutual conversions. For example, a  $\gamma$ -quantum is converted, under certain conditions, into an electron-positron pair according to the following reaction:

$$\gamma \rightarrow e^- + e^+. \quad (25.1.1)$$

This reaction cannot be interpreted in the sense that a  $\gamma$ -quantum consists of an electron and a positron. If it were so, an electron and positron had to combine into a single  $\gamma$ -quantum. In actual practice, two  $\gamma$ -quanta are emitted during the annihilation of an electron and positron. Consequently, reaction (25.1.1) cannot be treated as the decomposition of a compound particle into its constituents, but is rather the conversion of an elementary particle ( $\gamma$ -quantum) into other particles (electron and positron).

Another example of mutual transformation of elementary particles is the  $\beta$ -decay of a neutron (see Sec. 24.4):

$$n \rightarrow p + e^- + \nu. \quad (25.1.2)$$

A neutron spontaneously transforms into a proton ( $p$ ), electron ( $e^-$ ) and neutrino ( $\nu$ ). In many nuclei, a reverse process of conversion of a proton into a neutron, positron and neutrino<sup>2</sup> is observed:

$$p \rightarrow n + e^+ + \nu. \quad (25.1.3)$$

Since the proton transforms into the neutron and the neutron into the proton, this means that the two particles are equally elementary. As was mentioned in Sec. 24.8 in connection with atomic nuclei, electrons and positrons

---

<sup>2</sup> The rest mass of the proton is less than the sum of the rest masses of the neutron and the positron (the rest mass of the neutrino can be neglected, see Sec. 25.2). Consequently, according to Einstein's law, the rest energy of the products of reaction (25.1.3) is higher than the rest energy of the proton. Therefore, reaction (25.1.3) can proceed only provided that an energy is supplied from the outside. This condition is realized when a proton is a component of an unstable nucleus.

are formed in the process of the  $\beta$ -decay and are not present in the neutron or the proton.

The ability of elementary particles to undergo mutual transformations points to the complexity of their internal structure. This conclusion is also confirmed by other facts. By way of an example, let us consider the magnetic moment of the neutron.

Experiments show that like many other particles, the electron, the proton and the neutron are miniature tops. In other words, they possess a spin. Although spin is a property similar to the rotation about an axis passing through the centre of mass of a particle, there is no complete analogy here since the spin rotation cannot be accelerated or decelerated. The peculiarity of the behaviour of spin is determined by the wave nature of particles.

The rotation of an electrically charged particle leads to a circular current which imparts the properties of a small magnet to it (see Vol. 2). Indeed, the magnetism of electrons and protons has numerous manifestations. For example, the ability of iron to be magnetized, which is widely used in engineering, is a consequence of the spin magnetic moment of electrons. The magnetic moment of protons is much weaker, but it has also found some practical applications.

The fact that neutrons also exhibit the properties of small magnets seems more astonishing. It means that a neutron contains electric charges. Since neutron is neutral as a whole, the algebraic sum of its positive and negative charges must be equal to zero. But if unlike charges are at different distances from the rotational axis, the magnetic fields produced due to their motion will not be mutually compensated, and the neutron will be magnetized.

## 25.2. Neutrino

The *neutrino* is a neutral elementary particle that is emitted together with an electron or a positron during the  $\beta$ -decay of an atomic nucleus.

Like a neutron which is also deprived of an electric charge, a neutrino practically does not interact with electrons and does not cause a noticeable ionization of the medium. However, neutrons can easily be detected from their action on atomic nuclei (nuclear reactions and energy transfer during collisions, see Sec. 24.4), while neutrinos interact with nuclei very weakly. Until recently, no nuclear reactions initiated by neutrinos have been detected experimentally.

Then how the existence of neutrinos was confirmed if the particle is so elusive?

If only electrons were emitted during the  $\beta$ -decay, the energy of all  $\beta$ -electrons would be the same for a given radioactive isotope. Indeed, this energy would be equal to the difference in the internal energies of the initial

atomic nucleus and of the product nucleus + electron.<sup>3</sup> This difference should be the same since it was proved experimentally that all nuclei of a given isotope have the same mass, and hence the same internal energy. In actual practice, however, the energy of a  $\beta$ -electron can acquire various values from zero to a certain maximum value  $W_0$ . It is important that the maximum value is exactly equal to the internal energy liberated during the decay considered above. In order to correlate with the energy conservation law, we must assume that one more particle, viz. neutrino, is formed during the  $\beta$ -decay together with an electron. It is this particle that carries away the energy supplementing the electron energy to  $W_0$ . If the neutrino carries away an energy close to  $W_0$ , the energy of the electron is close to zero. If the energy of the neutrino is low, the electron energy, on the contrary, is close to  $W_0$ , and so on.

A detailed analysis of the  $\beta$ -decay has led to other not less convincing proofs of the emission of a neutrino in this process and made it possible to estimate the rest mass of the neutrino. It was found to be less than  $10^{-4}$  of the electron rest mass.

Many years of research resulted in 1956 in the experimental discovery of a nuclear reaction in which a neutrino ( $\nu$ ) was absorbed by a proton which was then converted into a neutron and a positron:



The source of neutrinos in these experiments was a powerful nuclear reactor in which neutrinos were formed during the  $\beta$ -decay of uranium fission fragments. Later, other reactions initiated by neutrino were observed in reactors (see Sec. 25.4).

Of great interest are the experiments on detecting the so-called solar neutrinos. These experiments make it possible to verify the validity of prevailing ideas about the solar structure and the nuclear processes occurring in the bulk of the Sun. The reaction of fusion of four protons, which is believed to be a source of solar energy (see Sec. 24.9), is accompanied by the emission of two neutrinos per helium nucleus formed. The neutrinos interact with a substance so weakly that their overwhelming majority pierce the Sun and escape into cosmic space.

A certain fraction of neutrinos reaching the Earth manifest themselves in that they cause nuclear reactions in a special detector. Since interactions involving neutrinos are very weak, this fraction is very small, the experiments on detecting solar neutrinos are expensive and complicated. However, they have been realized, and neutrinos emitted from the bulk of the Sun have been registered.

<sup>3</sup> The kinetic energy supplied during the  $\beta$ -decay to a heavy product nucleus is negligibly low in comparison with the energy of the electron.



### 25.3. Nuclear Forces. Mesons

In Sec. 24.8, we introduced the concept of nuclear forces, i.e. special type of forces acting between the particles constituting atomic nuclei, viz. neutrons and protons. Experiments (for example, on the investigation of nuclear reactions caused by fast neutrons and protons) have led to the conclusion that the nuclear forces of interaction between pairs of particles like proton-proton, neutron-proton or neutron-neutron are identical. Neutrons and protons behave in the same way in reactions which are determined only by nuclear forces. The difference in the properties of the neutron and the proton lies in a slightly larger mass of the former and in the electric charge of the latter and does not play a significant role in such phenomena. In order to emphasize the identity of the properties of neutrons and protons as regards nuclear forces, these two particles are combined by a common term *nucleons*.<sup>4</sup> By a nucleon we mean a neutron or a proton.

The most typical feature of nuclear forces is their *short range*. They attain very high values when nucleons are brought together to a distance of the order of  $10^{-13}$  cm. If, however, this distance is increased just several times, these forces decrease to such an extent that they can be neglected. In this respect, nuclear forces do not resemble electric forces or gravitational forces which vary smoothly (in inverse proportion to the squared distance between the particles). They rather resemble the forces emerging when two elastic balls come in contact.

The potential energy of the electric intersection of two protons can be calculated by formula (22.13.1) taken with the opposite sign (protons repel each other!). For the distance  $r$  between the protons equal to  $1.4 \times 10^{-13}$  cm, we have ( $e = 1.6 \times 10^{-10}$  C)

$$W = \frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \approx 1 \text{ MeV.}$$

Experiments show that the potential energy of the nuclear interaction between two nucleons brought to such a distance is about 50 MeV (if we assume that it is equal to zero at an infinitely large distance).

Thus, at small distances the nuclear interaction is about two orders of magnitude ( $10^2$  times) stronger than the electric interaction. At large distances between the protons like in an  $H_2$  molecule ( $r \approx 10^{-8}$  cm), the inverse relation takes place: the nuclear interaction between protons turns out to be negligibly weak as compared to the electromagnetic interaction.

How is nuclear interaction transferred from one nucleon to another?

The Einstein's theory of relativity states that no interaction between particles can be transferred at a velocity higher than the speed of light in vacu-

---

<sup>4</sup> A nucleon means a nuclear particle.

um. Let us suppose that one of two interacting particles has rapidly changed its position. The other particle will “feel” the change in the force of interaction due to the displacement of the first particle with a time lag proportional to the separation between the particles. The time lag can be much longer than the duration of the displacement of a particle: the particle may have come to a halt long ago, but something still occurs in the surrounding space, that will later affect the second particle. This means that there exists a certain agent transferring the interaction. In the case of interaction of charged particles, the agent is well known to us: it is the electromagnetic field. Besides the fields associated with electric charges, there also exist free electromagnetic fields, i.e. propagating electromagnetic waves (radiowaves, light, X-rays and  $\gamma$ -rays). It is well known that such free fields are the flows of electromagnetic quanta, viz. photons.

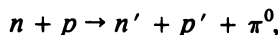
Similarly, other types of interaction (universal gravitation or nuclear interaction) have their own fields: *gravitational field* and the *field of nuclear forces*.

A nucleon produces a field of nuclear forces in the surrounding space, and this field acts on other nucleons getting into it. As was mentioned above, the range of the strong interaction is very small:  $10^{-13}$ - $10^{-12}$  cm.

In 1935, the Japanese physicist Hideki Yukawa (1907-1981) suggested that the nuclear field, like the electromagnetic field, can be not only bound but also free, i.e. there exist the *nuclear field quanta*. He proved that the short range of the nuclear field is due to the fact that the quanta of this field have nonzero rest masses. The larger the rest mass, the shorter the range of the forces. The observed range of nuclear forces is of the order of  $10^{-13}$  cm, which means that the rest masses of the quanta are 200-300 times as large as the rest mass of the electron.

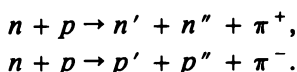
About ten years after the Yukawa prediction, new particles were discovered during the investigation of cosmic rays (see Sec. 25.8). The particles were called *pi-mesons* ( $\pi$ ). Subsequent investigations proved that these particles are just quanta of the nuclear field<sup>5</sup>. There are three types of  $\pi$ -mesons differing in electric charge: positive  $\pi$ -mesons ( $\pi^+$ ), neutral  $\pi$ -mesons ( $\pi^0$ ) and negative  $\pi$ -mesons ( $\pi^-$ ). The rest masses of  $\pi^0$ ,  $\pi^+$  and  $\pi^-$ -mesons are close and amount to about 270 rest masses of the electron. Just like the electromagnetic quanta emitted during braking of charges, nuclear quanta ( $\pi$ -mesons) are emitted during braking of nucleons, i.e. during their collisions.

Let us consider the simplest examples of reactions in which  $\pi$ -mesons are created:




---

<sup>5</sup> A meson (from the Latin word meaning intermediate) is a particle whose mass is intermediate between the electron mass and the nucleon mass.



The symbols  $n$  and  $p$  here stand for a neutron and a proton,  $n$ ,  $n'$ ,  $n''$ ,  $p$ ,  $p'$  and  $p''$  denote the nucleons differing in the state of motion (in the direction and magnitude of velocity). These reactions, like all known physical processes, are governed by the law of charge conservation.

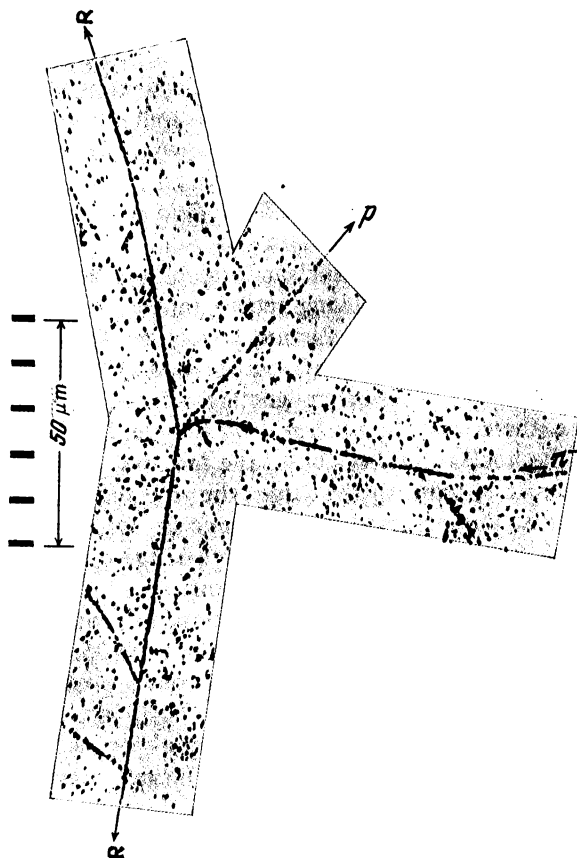


Fig. 414.

Splitting of a carbon nucleus as a result of capture of a  $\pi^-$ -meson is shown in the microphotograph of tracks of particles left in a photographic emulsion (see Sec. 25.6). The  $\pi^-$ -meson decelerated in the emulsion is attracted by a positive atomic nucleus and captured by one of its protons ( $\pi^- + p \rightarrow n$ ). This is accompanied by the liberation of a considerable amount of energy ( $\sim 138$  MeV), and the nucleus is split. The photograph shows the charged fragments, i.e. fast  $\alpha$ -particles and a proton; neutrons have left no track. The general reaction equation is as follows:  ${}^{12}_6\text{C} + \pi^- \rightarrow {}^4_2\text{He} + p + 3n$ .



Fig. 415.

Two cases of the decay of a  $\pi$ -meson are recorded in the photographic emulsion. A  $\pi$ -meson comes to a halt and decays according to the scheme  $\pi \rightarrow \mu + \nu$  (the left part of the photograph). The neutrino leaves no track. The formed muon also comes to a halt after traversing a distance of about 0.6 mm and decays according to the scheme  $\mu \rightarrow e + 2\nu$ . Fast particles ionize the medium to a smaller extent and form less dense tracks (compare the track of the electron and the initial part of the muon track). The charges of particles cannot be determined in this experiment, but in all probability they are positive since  $\pi^-$ -mesons are usually absorbed by nuclei in photographic emulsions before they have time to decay. On the contrary,  $\pi^+$ -mesons are repelled by nuclei, and it remains for them just to decay after the deceleration.

According to Einstein's law, the emission of a  $\pi$ -meson requires an energy not less than the rest energy of the  $\pi$ -meson, which is  $mc^2 \approx 140$  MeV. For this reason, a  $\pi$ -meson can be born only during collisions of particles having a rather high energy. In analogy with light quanta,  $\pi$ -mesons can be absorbed by nucleons to which they supply their kinetic energy, rest energy and the electric charge (Fig. 414).

Pi-mesons are unstable. A neutral  $\pi$ -meson decays in about  $10^{-16}$  s into two  $\gamma$ -quanta. Positive and negative  $\pi$ -mesons are transformed on the average in 30 ns ( $30 \times 10^{-9}$  s) into a positive  $\mu$ -meson (denoted by  $\mu^+$ ) and a neutrino and a negative  $\mu$ -meson ( $\mu^-$ ) and a neutrino respectively.

Mu-mesons (muons) are the particles with a rest mass equal to 207 rest masses of the electron and an average lifetime of  $2 \mu\text{s}$  ( $2 \times 10^{-6}$  s). Muons are converted into an electron or a positron and two neutrinos (Fig. 415).

Muons had been discovered before  $\pi$ -mesons and were initially regarded as nuclear quanta. This hypothesis was soon refuted since it turned out that muons interact with nucleons very weakly.

Several other types of still heavier and less stable mesons were discovered after  $\pi$ -mesons, which also strongly interacted with nuclei. Like  $\pi$ -mesons, they are also the quanta of the field of nuclear forces. Thus, the nuclear field is rather complicated and the complete theory of this field has not been created so far.

## 25.4. Particles and Antiparticles

At the end of the twenties, the newly developed quantum mechanics (see Sec. 22.17) together with the theory of relativity (see Sec. 22.6) was applied to explain the properties of the electron. This led to an unexpected conclusion: there should exist a positively charged twin of the electron! Indeed, after a few years such a particle was discovered (it is the well-known positron). The discovery of the positron was a triumph of the modern physical theory.

The positron is termed an *antiparticle* of the electron. The particle (electron) and the antiparticle (positron) differ only in the sign of the electric charge. Their other properties, i.e. the rest mass, the magnitude of charge and spin (i.e. internal rotation, see Sec. 25.1) are identical. The further evolution of quantum mechanics has led to the conclusion that *each particle*, except for a few neutral particles like photons and  $\pi^0$ -mesons, *must have an oppositely charged twin, viz. an antiparticle*.

In the previous section, we considered the two pairs of such twins:  $\pi^+$ - and  $\pi^-$ -mesons and muons  $\mu^+$  and  $\mu^-$ . Experiments show that like in the electron-positron pair, a particle and its antiparticle in each pair have similar properties, i.e. their masses and half-lives are equal.

The theory predicts that nucleons also have antiparticles, viz. the an-

tiproton and the antineutron (antinucleons). It is not surprising that the neutron whose total electric charge is equal to zero has an antiparticle that differs from it. Indeed, it was proved earlier that the neutron cannot be regarded a neutral particle. It has a complex internal charge distribution, which is manifested, in particular, in its nonzero magnetic moment. The magnetic moments of the neutron and of the antineutron turn out to have opposite directions relative to their spins.

In addition to the electric charge and magnetic moment, nucleons have one more important internal characteristic (quantum number) that distinguishes them from antinucleons. The existence of this characteristic, which can conditionally be termed a certain "charge", viz. *baryon charge*  $B$ , follows if only from the stability of nucleons. Indeed, in spite of their large mass, nucleons do not decay rapidly into light particles (electrons,  $\gamma$ -quanta or  $\pi$ -mesons) although such decays seem to be feasible from energy considerations. The stability of nucleons suggests that they have a certain quantum number which is conserved. This property was called the baryon charge. It does not exist for light particles, for this reason, the decay of nucleons into light particles was found to be forbidden.

Nucleons are ascribed a baryon charge  $B = +1$ . Then the antinucleons will have a baryon charge  $B = -1$ .

Thus, the antiproton is characterized by an electric charge of  $-1$  (in the units of elementary charge) and by a baryon charge  $B = -1$ . The electric charge of the antineutron is zero, while the baryon charge  $B = -1$ . The antiproton, like the proton, must be stable and have the same mass. The mass of the antineutron must be equal to the mass of the neutron, and the antineutron, like the neutron, must be unstable: it must be transformed into an antiproton through the  $\beta$ -transition.

Under terrestrial conditions, antinucleons should not exist for a long time since, like positrons, they annihilate, combining with nucleons and being transformed, as a rule, to the nuclear field quanta, viz.  $\pi$ -mesons.

Experiments show that like the electric charge, *the total baryon charge is conserved in all transformations of particle*. Therefore, on account of both charges, an antinucleon may be formed in nuclear reactions only in pair with a nucleon. Such reactions can be caused by particles having an energy of the order of billion electronvolts, which exceeds the rest energy of a nucleon-antinucleon pair (see Ex. IV.58).

In 1955-56, a few years after the first 6-GeV accelerator has been put in operation, a group of American physicists succeeded in detecting the processes of formation of antiprotons and antineutrons in experiments. These experiments not only reliably proved the existence of antinucleons but also confirmed the predictions of the theory concerning their properties. Figures 416 and 417 illustrate the investigation of antinucleons with the help of a bubble chamber (see Sec. 25.6).

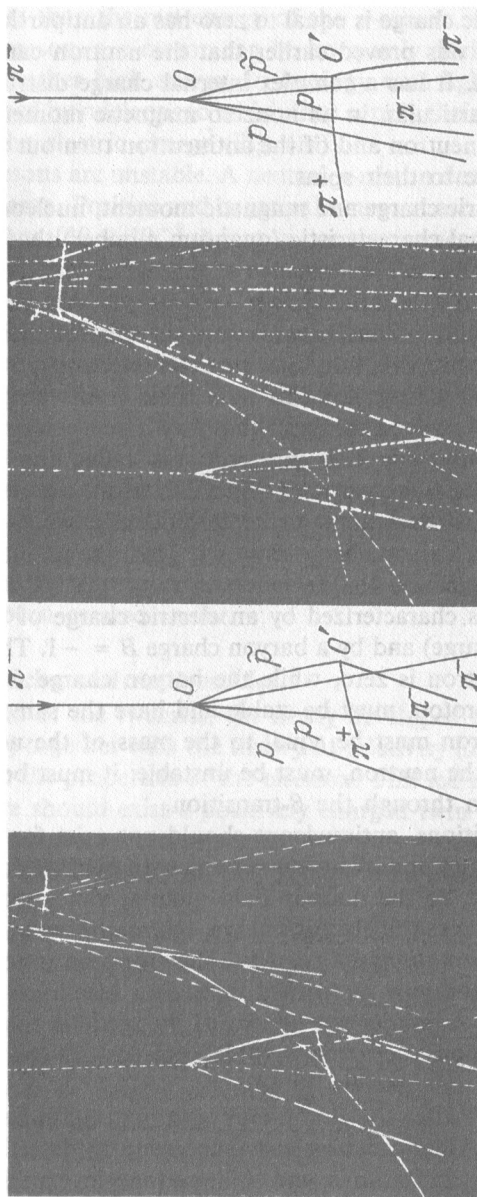


Fig. 416.

Formation and annihilation of an antiproton. The stereophotograph of the tracks in a bubble chamber filled with liquid propane  $C_3H_8$ . The chamber was irradiated to a beam of  $\pi$ -mesons and produced by the 10-GeV proton accelerator in Dubna having an energy of 7 GeV (the  $\pi$ -mesons were formed during the interactions of protons in beryllium). At point  $O$  (see the schematic diagram on the right of the photograph), a  $\pi$ -meson colliding with a proton forms a proton ( $p$ )-antiproton ( $\bar{p}$ ) pair (the reaction  $\pi^- + p \rightarrow \pi^- + p + \bar{p}$ ). The chamber was in a magnetic field. The curvature of the track of  $\bar{p}$  indicates that this particle bears a negative charge. At point  $O'$ , the antiproton collides with a proton and annihilates, giving birth to  $\pi^+$ - and  $\pi^-$ -mesons, and, as follows from the analysis of the photographs by taking into account the energy and momentum conservation laws, to  $\pi^0$ -meson which does not leave any track in the chamber.



**Fig. 417.**

Formation and annihilation of an antineutron. A propane bubble chamber was irradiated by a beam of antiprotons formed during the collisions of protons having an energy of 6 GeV with a beryllium target. The track of one of the antiprotons abruptly terminates (upper arrow), while other antiprotons having the same energy cross the entire chamber. This can be explained only by the fact that the reaction  $\bar{p} + p \rightarrow \bar{n} + n$  has occurred. The antineutron and the neutron fly in the directions close to the direction of the antiproton since it has transferred its momentum to these particles, without leaving any track in the chamber. At the point indicated by the lower arrow (it lies approximately on the continuation of the track of the antiproton), the antineutron collides with a proton or a nucleus. Charged products of annihilation (mainly  $\pi$ -mesons) form a "star" on the photograph. From the bending of tracks in the magnetic field, we can say that both positive and negative particles are emitted.



Later, *antideuterons* (viz. atomic nuclei consisting of an antiproton and an antineutron) were detected among the products of nuclear reactions between high-energy particles. Theoretically, various nuclei (or, to be more precise, *antinuclei*) can be constructed from antiprotons and antineutrons, which would differ from normal proton-neutron nuclei only in the sign of the electric (and baryon) charge<sup>6</sup>. Combining with positrons, such antinuclei should form atoms as stable as normal terrestrial atoms. This means that an *antimatter* formed by antinucleons and antielectrons (positrons) may exist.

Astrophysical observations have not resulted so far in the discovery of a significant amount of antimatter in the visible part of the Universe. It is difficult to say for sure whether this is a result of the insufficient accuracy of observations or the Universe is indeed asymmetric, i.e. is constructed only by matter although the antimatter is an equally good building material.

In the previous discussion, we considered the neutrino as a single particle. However, recent works have proved that there are several types of neutrino.<sup>7</sup> The  $\beta$ -decay of neutrons and protons results in the formation of electrons  $e^-$  and positrons  $e^+$ . The particle emitted together with an electron was agreed to be called the *electron antineutrino*  $\bar{\nu}_e$ . Then the particle emitted together with a positron should be called the *electron neutrino*  $\nu_e$ . Therefore, the  $\beta$ -decay reactions (25.1.2) and (25.1.3) can be written as follows:

$$n \rightarrow p + e^- + \bar{\nu}_e, \quad (25.4.1)$$

$$p \rightarrow n + e^+ + \nu_e. \quad (25.4.2)$$

By adding  $\bar{\nu}_e$  to the right- and left-hand sides of (25.4.2) and annihilating neutrino  $\nu_e$  and antineutrino  $\bar{\nu}_e$  on the right-hand side (the liberated energy is absorbed by the positron), we arrive at (25.2.1)<sup>8</sup> in the refined form:

$$\bar{\nu}_e + p \rightarrow n + e^+. \quad (25.4.3)$$

Similarly, it follows from (25.4.1) that

$$\nu_e + n \rightarrow p + e^-. \quad (25.4.4)$$

Are neutrino  $\nu_e$  and antineutrino  $\bar{\nu}_e$  identical or different particles? The answer to this question must be given by experiment. We have already considered the particles having zero electric charge and differing from their antiparticles. These are neutrons and antineutrons differing in the sign of baryon charge. However, there are neutral particles of another type which are identical to their antiparticles, like photons or  $\pi^0$ -mesons that are referred to as *truly neutral particles*. Experiments carried out with the beams of antineutrinos produced

<sup>6</sup> In 1970, the nuclei of antihelium-3, i.e. the nuclei consisting of two antiprotons and an antineutron were synthesized at the Institute of High-Energy Physics in Serpukhov. Later, the nuclei of antitritium-3 consisting of an antiproton and two antineutrons were also obtained.

<sup>7</sup> For detail, see Sec. 26.5.

<sup>8</sup> This analysis is of the general nature. It can be used to show that *any* particle can be transformed from one side of a reaction equation to the other by replacing it by its antiparticle.

in a nuclear reactor<sup>9</sup> showed that reaction (25.4.3) of absorption of  $\bar{\nu}_e$  by protons is observed indeed (see Sec. 25.2). However, the absorption of  $\bar{\nu}_e$  by neutrons has never been observed. But this is not surprising if we assume that the electron neutrino and antineutrino are different particles (in this case,  $\nu_e$  and not  $\bar{\nu}_e$  can be absorbed during interactions with neutrons!). Thus, direct experiments indicate that *the electron neutrino  $\nu_e$  and antineutrino  $\bar{\nu}_e$  are different particles*, and for this reason they are not truly neutral. Further investigations revealed that neutrinos formed during a decay of  $\pi$ -mesons together with muons differ from those formed during  $\beta$ -decays (25.4.1) and (25.4.2) together with electrons.

The reaction for the decay of a  $\pi^+$ -meson into a muon and a neutrino should be written as  $\pi^+ \rightarrow \mu^+ + \nu_\mu$ .<sup>10</sup> By adding  $\mu^-$  and a neutron on the right- and left-hand sides, annihilating  $\mu^+$  and  $\mu^-$  and combining  $n$  and  $\pi^+$  into  $p$ , we obtain

$$\mu^- + p \rightarrow n + \nu_\mu.$$

Obviously, the inverse reaction must also take place:

$$\nu_\mu + n \rightarrow p + \mu^-.$$

This reaction was observed in experiments with accelerators operating on the beams of neutrinos  $\nu_\mu$  formed during the decay of  $\pi^+$ -mesons. However, these beams did not cause reactions (25.4.3) and (25.4.4). Hence, it was concluded that the muon neutrino and the electron neutrino are different.

It was also shown experimentally that the muon neutrino and antineutrino ( $\nu_\mu$  and  $\bar{\nu}_\mu$ ) are different. Various types of neutrino are considered in greater detail in Sec. 26.5.

## 25.5. Particles and Interactions

It is believed now that the entire variety of phenomena occurring in the Universe at all levels (i.e. in the microcosm, on the Earth, stars and in galaxies) is determined just by *four* types of interaction. Two of them are known from classical physics. These are the *gravitation* (universal gravitation) and the *electromagnetic interaction*. Two other types of interaction, viz. the *nuclear* interaction (which is often referred to as *strong*) and the *weak* interaction have a short range and are hence not manifested directly either in the motion of macroscopic bodies or in the properties of atoms and molecules. The latter types of interaction are exhibited only in nuclear phenomena and in transformations of elementary particles. The strong interaction was considered in Sec. 25.3. The *weak interaction* is a *special* type of interaction involved in all processes with neutrinos, like the capture of a neutrino by nuclei, the  $\beta$ -decay, and the decay of  $\pi^+$ -,  $\pi^-$ -mesons and muons.

The force of interaction between two particles can be characterized by the potential energy of their approach to a certain distance. Let us compare the energies of strong, electromagnetic,

<sup>9</sup> In nuclear reactors, the  $\beta^-$ -decay of uranium fission products "overloaded" with neutrons takes place (i.e. the  $\beta^-$ -decay of neutrons). For this reason, the reactors are high-intensity sources of antineutrinos  $\bar{\nu}_e$ .

<sup>10</sup> The particle  $\nu_\mu$  is referred to as the muon neutrino. The *muon antineutrino*  $\bar{\nu}_\mu$  is formed during the decay  $\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$ .

weak and gravitational interactions between two protons at a distance  $r \approx 10^{-13}$  cm, where the strong interactions are exhibited practically on full scale. In Sec. 25.3, we compared the corresponding estimates of the energy of the electric ( $\sim 1$  MeV) and the strong ( $\sim 50$  MeV) interactions between these particles. The energy of the weak interactions between these particles is of the order of  $10^{-6}$  eV. In order to calculate the potential energy of the gravitational interaction between the protons, we shall use formula (5.16.10) of Vol. 1:  $W = -Gm_p^2/r$  ( $G = 6.7 \times 10^{-11}$  N·m<sup>2</sup>/kg<sup>2</sup> is the gravitational constant,  $m_p = 1.67 \times 10^{-27}$  kg is the mass of the proton,  $r = 10^{-15}$  m is the separation between the protons which are regarded as point masses). Then,

$$|W| = G \frac{m_p^2}{r} = 6.7 \times 10^{-11} \frac{(1.67 \times 10^{-27})^2}{10^{-15}} \text{ J} \\ = 1.87 \times 10^{-49} \text{ J} = \frac{1.87 \times 10^{-49}}{1.6 \times 10^{-19}} \text{ eV} \approx 10^{-30} \text{ eV}.$$

This energy is extremely low, and no significant manifestation of gravitation in the microcosm has been discovered so far.

Thus, the energies of fundamental interactions are to one another as follows: strong:electromagnetic:weak:gravitational =  $1:10^{-2}:10^{-14}:10^{-38}$ .

The concepts of time characteristic of a certain phenomenon play an important role in physics of elementary particles. Let us establish first of all the time scale for the processes involving strong interactions. In order to estimate this scale, let us consider nuclear collisions between fast particles (having a velocity comparable with the speed of light  $c$ ). Since the range of nuclear forces  $r \approx 10^{-12} - 10^{-13}$  cm, the time of such a strong interaction will be characterized by the quantity  $r/c \sim 10^{-22} - 10^{-23}$  s. This also means that if the decays of particles are due to strong interactions, the corresponding lifetime will be just of the order of magnitude of this small quantity (short-lived particles). If, for some reason or other, "strong decays" cannot occur and a particle decays under the action of electromagnetic forces, its lifetime will range between  $10^{-16}$  and  $10^{-20}$  s. The corresponding time for "weak decays" has the order of  $10^{-8} - 10^{-13}$  s. For this reason, the particles decaying only due to weak interactions are considered long-lived in the world of elementary particles.

Of the four known interactions (gravitational, weak, electromagnetic and strong), only the gravitational interaction is universal since all particles without any exception experience gravitational forces.

According to the type of interaction, in which they participate, particles are divided into several classes.

The first class includes only one particle, viz. photon.<sup>11</sup> A photon interacts (is emitted or absorbed) with electric charges, i.e. exhibits the electromagnetic interaction. The strong and weak interactions are *not* inherent in the photon.

<sup>11</sup> Some other particles like gluons intermediate bosons (see Sec. 26.5) have similar properties.

The second class incorporates the so-called *leptons*<sup>12</sup>, i.e. the electron, muon, neutrino and their antiparticles. The common property of leptons is that they exhibit the weak interaction and do not participate in the strong interactions. Charged leptons (electron and muon) are naturally subjected to the electromagnetic interaction.<sup>13</sup>

The third, largest class is formed by *hadrons*<sup>14</sup>, viz. strongly interacting particles. Hadrons exhibit all four types of interaction.

*Mesons*, viz. strongly interacting particles having zero *baryon* charge, form the first subgroup of hadrons. As was mentioned earlier, they should be considered the quanta of nuclear field (the field of strong interactions).

The second subgroup includes *baryons*, viz. particles having a *baryon* charge.

The lightest baryons are nucleons (neutrons and proton) which are stable (the neutron is stable in nuclei) and together with electrons serve as bricks of which the matter is constructed. Ultimately, this is due to the law of baryon charge conservation which allows a baryon to vanish only together with an antibaryon. The baryon charge conservation makes impossible, for example, the destruction of atoms through the annihilation of a proton and an electron (the conversion into  $\gamma$ -quanta or mesons). Since antibaryons practically are not available in our world, nucleons cannot vanish. In this respect, they strongly differ from photons and mesons which ultimately vanish (are absorbed or decay), transferring their energy (and electric charge in the case of charged mesons) to leptons or nucleons. Recently, hundreds of heavier and less stable mesons and baryons have been discovered. Certain regularities in their properties like the mass, mode of formation and decay have been established.<sup>15</sup> However, a consistent theory that would describe the properties of hadrons as successfully as the quantum theory describes atoms and molecules has not been created so far. A more fundamental question as to why particular elementary particles (electron, proton, photon, neutrino, etc.) have specific properties also remains unanswered.

## 25.6. Detectors of Elementary Particles

In Chap. 23 we described the devices for detecting microparticles (the Wilson cloud chamber, the scintillation counter and the gas-discharge counter). Although these detectors are often used for investigating elementary parti-

---

<sup>12</sup> The term "leptons" is derived from the Greek *leptos* meaning light.

<sup>13</sup> According to modern theories, the electromagnetic and weak interactions are different manifestations of a more general "electroweak" interaction.

<sup>14</sup> The term "hadrons" is derived from the Greek *hadros* meaning large, strong.

<sup>15</sup> Properties of hadrons are described in greater detail in Sec. 26.2.

cles, they are sometimes inconvenient. As a matter of fact, the most interesting interaction processes accompanied by mutual transformations of particles are rather rare events. Before an interesting collision takes place, a particle must meet on its way a large number of electrons or nucleons. In actual practice, it has to cover a distance of the order of ten centimetres or even metres in a dense substance (a charged particle having an energy of billions electronvolts loses over this distance only a fraction of its energy as a result of ionization).

However, the sensitive layer in the Wilson cloud chamber or in the gas-discharge counter is extremely thin (when recalculated for a dense substance). For this reason, some other methods for detecting particles have been worked out.

The photographic method proved to be very fruitful. A charged particle passing across a special fine-grained photographic emulsion leaves a track which after the development of the plate is seen in a microscope as a chain of black grains. The type of the particle (viz. its charge, mass and energy) can be determined from the shape of the track left by it in the photographic emulsion. The photographic method is convenient not only because thick layers of substances can be used, but since, unlike in the Wilson cloud chamber, the traces of charged particles do not disappear soon after a particle has passed through it. When rare events have to be investigated, photographic plates can be exposed for a long time. This is especially important for investigating cosmic rays. The examples of rare events recorded on a photographic emulsion are presented in Figs. 414 and 415. Figure 418 is even more interesting.

Another remarkable method is based on the properties of *superheated liquids* (see Vol. 1, Sec. 17.12). When a very pure liquid is heated to a temperature slightly exceeding the boiling point, the liquid does not boil since the surface tension prevents the formation of steam bubbles. The American physicist Donald Glaser (b. 1926) discovered in 1952 that a superheated liquid boils instantaneously when exposed to an intense  $\gamma$ -radiation: the additional energy liberated in the traces of fast electrons produced by the  $\gamma$ -radiation in the liquid facilitates the formation of bubbles in the liquid.

Glaser used this effect for designing the so-called liquid *bubble chamber*. A liquid is heated under an elevated pressure to a temperature close to but lower than the boiling point. Then the pressure, and hence the boiling point, is lowered, and the liquid turns out to be superheated. The trace of vapour bubbles is formed along the trajectory of a charged particle passing through the liquid at this moment. With an appropriate illumination, it can be photographed. Bubble chambers are placed as a rule between the poles of a strong electromagnet so that the magnetic field bends the trajectories of particles. Having measured the length of the trace left by a particle,

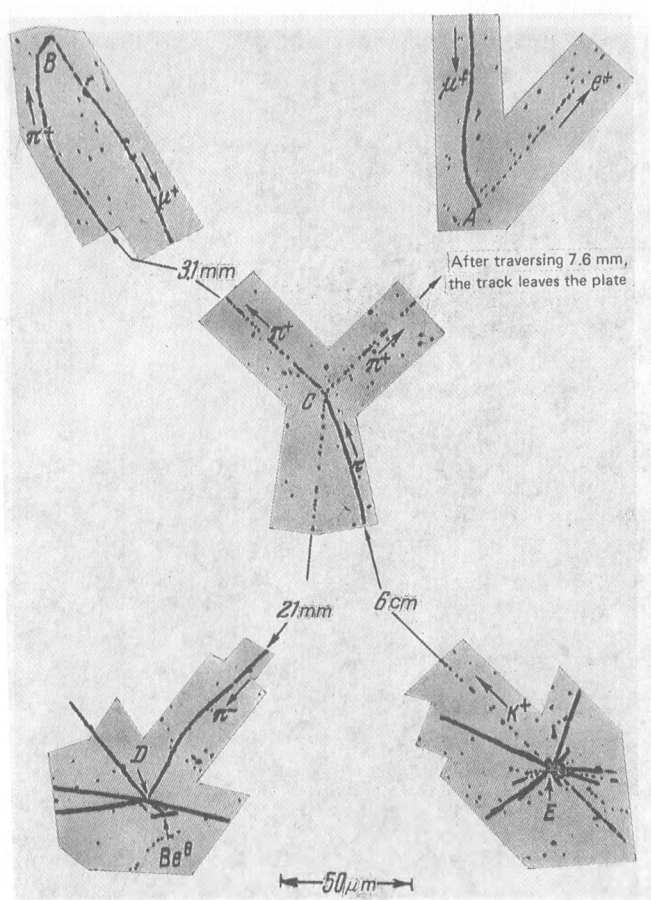
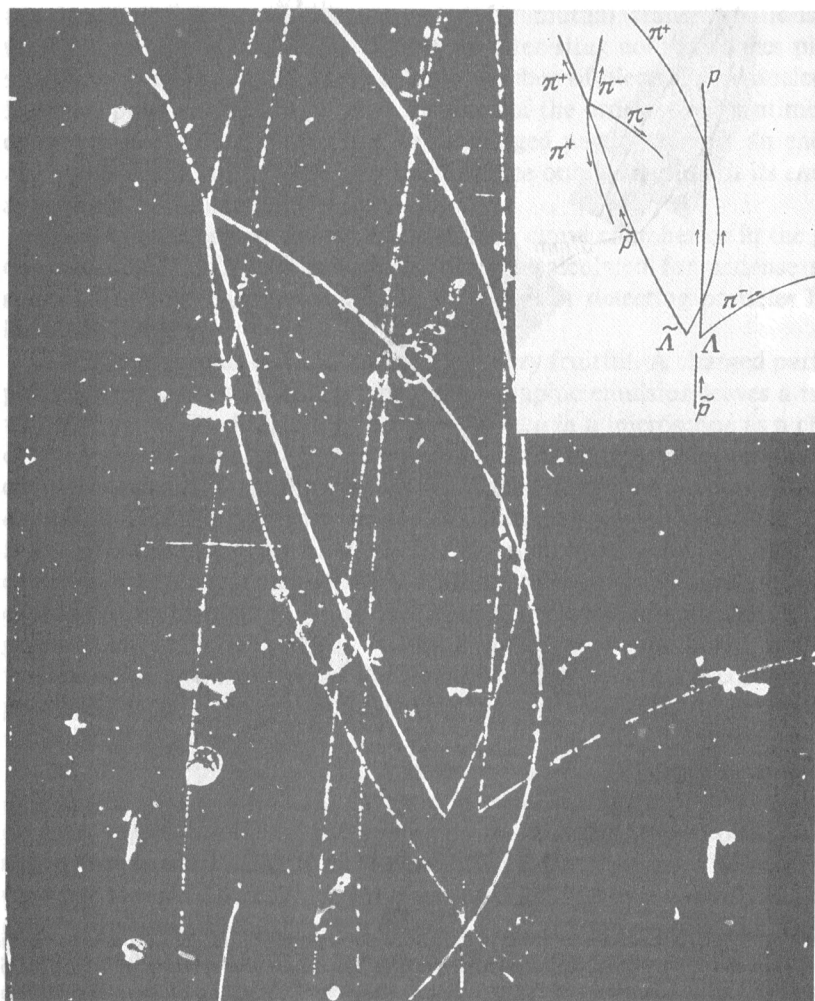


Fig. 418.

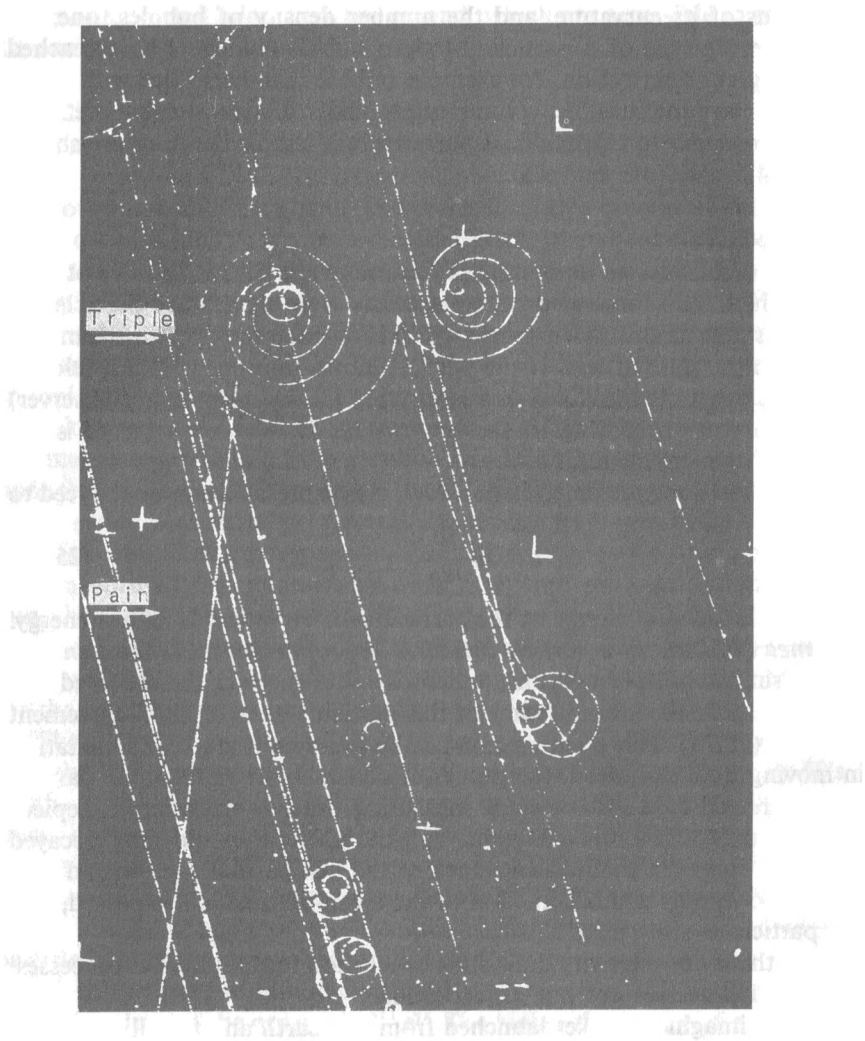
Transformations of particles recorded in the pile of photographic emulsions exposed to cosmic rays. At point *E*, an invisible fast neutral particle has caused the splitting of a nucleus of the photographic emulsion and formed mesons (the "star" consisting of 21 tracks). One of the mesons ( $K^+$ -meson), having covered a distance of 6 cm (the photograph shows only the beginning and the end of the track since for the magnification used here the length of the track would be 30 m), comes to a halt at point *C* and decays according to the scheme  $K^+ \rightarrow \pi^+ + \pi^+ + \pi^-$ . The  $\pi^-$ -meson whose track is directed downwards is captured by an  $^{16}\text{O}$  nucleus at point *D*, thus causing its splitting. One of the fragments is a  $^8\text{Li}$  nucleus which is converted into a  $^8\text{Be}$  nucleus as a result of the  $\beta$ -decay. The latter immediately splits into two  $\alpha$ -particles flying apart in the opposite directions (they form a "hammer" on the photograph). The  $\pi^+$ -meson comes to a halt and is converted into a  $\mu^+$ -meson (and a neutrino) at point *B*. The termination of the track of the  $\mu^+$ -meson is shown in the upper right corner of the photograph. We can also see the track of the positron formed during the decay

$$\mu^+ \rightarrow e^+ + 2\nu.$$



**Fig. 419.**

Formation and decay of  $\Lambda$ -hyperons. The reaction  $\bar{p} + p \rightarrow \bar{\Lambda} + \Lambda$  is registered in a hydrogen bubble chamber placed in a magnetic field and irradiated by antiprotons. The reaction takes place at the point of termination of the track of  $\bar{p}$  (see the schematic diagram in the upper part of the figure). The neutral  $\lambda$ - and anti- $\lambda$ -hyperons, having covered a small distance without leaving a track, decay according to the schemes  $\Lambda \rightarrow p + \pi^-$  and  $\bar{\Lambda} \rightarrow \bar{p} + \pi^+$ . The antiproton  $\bar{p}$  annihilates with a proton, giving birth to two  $\pi^+$ -mesons and two  $\pi^-$ -mesons.



**Fig. 420.**

Tracks of electron-positron pairs in a bubble chamber. The hydrogen chamber is irradiated by high-energy charged particles and  $\gamma$ -quanta. The triple of tracks (the arrow "triple") is a result of interaction between a  $\gamma$ -quantum and an electron, leading to the formation of an  $e^+e^-$ -pair (spirals wound in opposite directions). A weakly bent track belongs to the primary electron which has acquired a high energy in this process. The arrow "pair" indicates the  $e^+e^-$ -pair formed by the  $\gamma$ -quantum on a proton. The proton does not leave a visible track since it does not acquire a large energy during the interaction with the  $\gamma$ -quantum because of its large mass.



the radius of its curvature, and the number density of bubbles, one can determine the type of a particle. Modern bubble chambers have reached a high degree of perfection. For example, bubble chambers filled with liquid hydrogen have the sensitive volume of the order of several cubic metres<sup>16</sup>. The photographs of the tracks of particles in a bubble chamber are shown in Figs. 416, 417, 419 and 420.

### 25.7. Clock Paradox

Let us now consider an interesting prediction of the Einstein theory of relativity, which was confirmed in experiments with elementary particles.

Let us take an unstable particle which is characterized by the mean lifetime  $\tau_0$  in the state of rest. If the particle moves uniformly with a velocity  $v$ , its mean lifetime measured in a laboratory (i.e. by a stationary observer) should increase according to the law  $\tau = \tau_0 / \sqrt{1 - v^2/c^2}$ , where  $c$  is the speed of light in vacuum.

Using relations presented in Sec. 22.7, this expression can be reduced to

$$\tau = \tau_0 \frac{W}{W_0}, \quad (25.7.1)$$

where  $W$  is the total energy of the particle and  $W_0 = mc^2$  is its rest energy. *The mean lifetime of a particle increases in proportion to its total energy.* In experiments with fast muons,  $\pi$ -mesons and  $K$ -mesons, the observed ten-fold increase in the mean lifetime of this particle was in brilliant agreement with law (25.7.1). This phenomenon can be characterized as time dilatation in moving bodies. Indeed, the processes occurring in an unstable particle can be treated as a sort of *clock* measuring time. A stationary timepiece has measured a few mean lifetime, and the particle should have decayed long ago. However, the intrinsic clock of the fast particle lags behind. According to it, only a small fraction of the mean lifetime  $\tau_0$  has passed, and the particle is still "alive".

The theory of relativity draws this conclusion for all physical processes. Biological processes are not an exception.

Let us imagine a rocket launched from the Earth and travelling in cosmic space at a velocity close to the speed of light. At the moment the rocket returns to the Earth, the timepiece on it will indicate the shorter time interval than the clocks on the Earth. Astronauts will age less than their friends who stay on the Earth. The validity of these conclusions seem doubtful, and they are known as the clock paradox. Experiments with unstable particles, however, prove that the clock paradox is a scientific fact. It should be noted that the dilatation of time is negligibly small for flight velocities

---

<sup>16</sup> The largest available hydrogen bubble chamber has a volume of 30 m<sup>3</sup>.

of the order of ten kilometers per second, which are accessible to modern astronautics.

### 25.8. Cosmic Radiation (Cosmic Rays)

It was noted even in the first experiments on radioactivity that a very small current is observed in an ionization chamber (see Fig. 376) even in the absence of radioactive specimens. The presence of this current indicates that some sort of radiation causes ionization in the chamber, which is known as residual ionization. At first, residual ionization was explained by admixtures of radioactive substances in the soil and in the atmosphere. If it were true, residual ionization should be reduced as the ionization chamber moves away from the surface of the Earth. However, the experiments in which an ionization chamber was lifted on balloons to a high altitude led to the opposite result. The residual ionization at an altitude of 9 km turned out 40 times stronger than on the ground. This result becomes clear if we assume that the radiation responsible for residual ionization reaches the Earth from *outside*, and on its way through the atmosphere is gradually absorbed by it. Subsequent experiments confirmed the extraterrestrial origin of radiation and proved that its intensity weakly depends on the position of the Sun, the Moon and other celestial bodies on the sky. Hence it follows that the radiation is emitted not by an individual celestial body but by the outer space uniformly in all directions. For this reason, the radiation responsible for the residual ionization was termed *cosmic radiation*, or *cosmic rays*.

The nature of cosmic radiation turned out to be rather complex. The idea about this phenomenon as a whole has been formulated only by 50's on the basis of the results of numerous investigations among which the works of the school headed by the Soviet physicist D. V. Skobeltsyn occupy an important place. According to modern ideas, the primary cosmic radiation, i.e. the radiation arriving in the atmosphere from the deep of the outer space, mainly consists of fast positive particles (protons) and some  $\alpha$ -particles and other nuclei. The primary particles of cosmic rays have a huge energy of the order of billion electronvolts. Sometimes, this energy attains fantastic values of the order of  $10^{21}$  eV. The higher the energy of a particle, the smaller the number of such particles in the primary radiation component. There are several hypotheses about the mechanism of acceleration of particles in the Universe to such drastic energies. The investigations in this field are far from being completed.

Only a small fraction of the primary cosmic radiation reaches the surface of the Earth. The overwhelming part of primary particles collide with nuclei of the atoms constituting air in the upper layers of the atmosphere. In view of the huge energy of primary particles, such collisions result in the splitting of atomic nuclei accompanied by the emission of fast neutrons,

protons and  $\alpha$ -particles. Besides, the collisions of high-energy particles with nuclei are accompanied by the formation of *new* particles, viz. various types of mesons and hyperons (see Sec. 25.5). Depending on the type, hyperons are transformed into a meson and a nucleon (neutron or proton). Ultimately, mesons are transformed into positrons or  $\gamma$ -quanta.

Thus, a collision of a fast primary particle with an atomic nucleus results in the formation of a large number of secondary particles having a lower energy, i.e. protons, neutrons,  $\alpha$ -particles, various hyperons and mesons, electrons, positrons and  $\gamma$ -quanta. An example of such a process is shown in Fig. 418. Moving in the atmosphere, secondary particles in turn are multiplied at the expense of nuclear splitting and other processes like the formation of electron-positron pairs from  $\gamma$ -quanta (see Sec. 24.6).

Along with the multiplication of particles in the atmosphere, their absorption also takes place in the same way as the absorption of  $\alpha$ -,  $\beta$ - and  $\gamma$ -particles passing through a substance. The multiplication is the prevailing process in the upper layers of the atmosphere, and the number of particles in the cosmic radiation increases up to an altitude of about 20 km above the sea level. Below this level, absorption plays the major role, and the radiation intensity drops. The curve representing the intensity of cosmic radiation as a function of altitude is shown in Fig. 421.

The total energy carried by cosmic rays to the Earth is rather small as compared to the energy carried by luminous radiation of the Sun. For this reason, the effect of cosmic radiation on the inanimate objects on the Earth is obviously insignificant. In the evolution of life, it possibly has played a significant role since ionizing radiations increase the frequency of mutations, and hence the rate of evolution. The investigation of cosmic radiation is very important for the study of elementary particles and the Universe. Cosmic rays can be regarded as a natural laboratory in which the processes

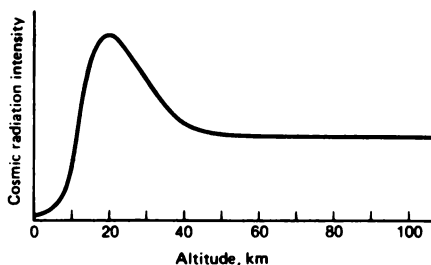


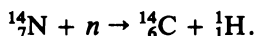
Fig. 421.

The intensity of cosmic radiation as a function of the altitude above the sea level. At an altitude above 50 km, only the primary component of cosmic rays arriving from the space is present, and the radiation intensity does not depend on altitude. Below 50 km, the radiation intensity first increases due to the formation of secondary particles and then decreases due to the increasing absorption in the atmosphere.

of interaction of high-energy particles whose energy exceeds the energy supplied by the most powerful laboratory accelerators to elementary particles occur. As the energy of elementary particles increases, the phenomena caused by them become more and more diversified, and the properties of particles are manifested on a larger scale.

The investigation of cosmic radiation has led to the discovery of the positron and a number of mesons. A more detailed study of these particles was carried out later with the help of accelerators. There are all grounds to believe that the investigation of cosmic rays will provide a valuable information about elementary particles especially now, when cosmic laboratories on satellites are used for this purpose. The role of cosmic radiation as a source of astrophysical information, i.e. the information about the processes occurring in far regions of the Universe where the radiation is generated and starts to propagate, becomes more and more important.

**Radiocarbon dating in archeology.** The neutrons of cosmic rays interacting with atmospheric nitrogen form a  $\beta$ -radioactive isotope of carbon  $^{14}\text{C}$ , which is known as radiocarbon (the half-life is 5730 years):



Radiocarbon, like the ordinary carbon  $^{12}\text{C}$ , is contained in air in the form of carbon dioxide, in the proportion  $^{14}\text{C}:^{12}\text{C} \approx 1:10^{12}$ . Since the chemical properties of all carbon isotopes are very close, the same proportion is retained in plants that absorb atmospheric carbon dioxide as well as in organisms of animals fed by plants. Thus, animals and plants have a weak but still noticeable radioactivity.

The absorption of carbon dioxide by an animal or a plant ceases with its death, and the activity of  $^{14}\text{C}$  in remnants gradually decreases (by half during each half-life, i.e. in every 5730 years). Comparing the radioactivity of living fossils (per 1 g of carbon) with the radioactivity of existing plants and animals, one can determine the degree of the decay of  $^{14}\text{C}$ , and hence the age of fossils.

The validity of this idea was verified in experiments on the measurement of radioactivity of the objects of known age, in particular, samples of wood extracted from the tombs of Egyptian pharaohs Zoser and Snofru. The measured activity of  $^{14}\text{C}$  was in good agreement with the known dates of deaths of these pharaohs (about 2700-2625 BC).

The results of similar experiments prove that the content of  $^{14}\text{C}$  per unit volume of atmospheric carbon dioxide remains unchanged during the last 50-100 thousand years, and that indeed no carbon metabolism occurs after the death of an organism. This formed the basis of the radiocarbon method of determining the age (dating), which is now extensively and successfully used in archeology.

- ? **IV.58.** Determine the minimum kinetic energy of protons required for the formation of (a)  $\pi^0$ -meson in the reaction  $p + p \rightarrow p + p + \pi^0$ , (b) a proton-antiproton pair in the reaction  $p + p \rightarrow p + p + \bar{p} + p$ .
- IV.59.** Knowing the mass of a neutral  $\pi$ -meson ( $135.0 \text{ MeV}/c^2$ ), determine the energy of  $\gamma$ -quanta formed during the decay of a stationary neutral  $\pi$ -meson:  $\pi^0 \rightarrow 2\gamma$ .
- IV.60.** Determine the maximum energy of electrons emitted during the  $\beta$ -decay of a neutron if the neutron mass is  $939.57 \text{ MeV}/c^2$ , and the mass of the hydrogen atom is  $938.73 \text{ MeV}/c^2$ .

## Chapter 26

# New Achievements in Elementary-Particle Physics

### 26.1. Accelerators and Experimental Technology

During last decades, a real revolution took place in elementary-particle physics, which considerably changed our views on the nature of matter. This revolution was primarily due to a rapid development of accelerators and experimental technology. An increase in the energy of accelerators which remain the main tool of investigating elementary particles plays an important role for several reasons.

1. An increase in energy makes it possible to obtain new types of elementary particles with large masses. At lower energies, such particles just cannot be formed in view of the energy and momentum conservation laws (see Ex. IV.58 for the energy threshold).

2. Accelerators can be compared with giant microscopes that make it possible to investigate space over very short distances comparable with the de Broglie wavelength for accelerated particles. For example, the particles having an energy of  $1 \text{ TeV} = 10^3 \text{ GeV}$  are characterized by the de Broglie wavelength  $\lambda = h/p \approx 1 \times 10^{-16} \text{ cm}$ .<sup>1</sup> Such particles can be used for probing the regions of space as small as  $10^{-16} \text{ cm}$  where some new regularities of the physics of microcosm, which could not be noticed over large distances, may be manifested.

3. An increase in the energy of particles is accompanied by the change in the nature of interaction between them, as well as in the characteristic of the known processes. It may turn out that some features of these phenomena are manifested more clearly at high energies. For example, the common nature of weak and electromagnetic interaction was established just in experiments with high-energy particles.

In recent years, accelerators have been designed that look giant even in comparison with a huge Serpukhov accelerator (see Fig. 393). They allowed the energy required for the formation of new particles to be increased by

---

<sup>1</sup> In the selected units (see the footnote on p. 498),  $p = 1 \text{ TeV}/c$ , and

$$\lambda = \frac{h}{p} = \frac{4.14 \times 10^{-24} \times 3 \times 10^{10} \text{ GeV} \cdot \text{cm}}{10^3 \text{ GeV}} = 1.24 \times 10^{-16} \text{ cm}.$$

about two orders of magnitude. The experiments on colliding-beam accelerators with storage rings played an important role.

What is the difference between these accelerators?

In experiments with “ordinary” accelerators, or fixed-target accelerators, the processes of interaction of accelerated particles with “stationary targets” (i.e. nucleons and nuclei of the atoms of the target material) are being investigated. In this case, only a small fraction of the energy of accelerated protons or electrons can be spent “usefully” (to form new particles). Since the particles bombarding the target have a very large initial momentum, according to the momentum conservation law, this momentum must be carried away by all the secondary particles formed during the interaction. Naturally, these particles will have a rather high kinetic energy. Thus, the large part of the initial energy is converted into the kinetic energy of the products of a nuclear reaction, and only a comparatively small part of this energy can be spent on the formation of new particles.

Let us consider again the solution of Ex. IV.58 in which it was shown that for the formation of a proton-antiproton pair in the reaction  $p + p \rightarrow p + p + p + \bar{p}$ , the primary proton must have a kinetic energy  $W_k > 6mc^2$ , although the “useful energy expenditures” amount to only  $2mc^2$ . The remaining energy is transformed into the kinetic energy of secondary particles. A similar situation is observed in other processes as well.

Unlike fixed-target accelerators, colliding-beam accelerators with storage rings make it possible to utilize the whole of the initial energy. The operating principle of these accelerators mainly consists in that two very intense and well-focussed beams of accelerated particles are produced and directed against each other so that a head-on collision between them occurs. In this case, the total momentum of two colliding particles is zero, and the formed secondary particles may have a very low kinetic energy (the generation threshold corresponds to the formation of stationary particles). For example, a head-on collisions of two protons having a kinetic energy  $W_k \geq mc^2$  may result in the formation of a proton-antiproton pair, and we have a considerable gain in energy.

Some years ago, experiments with colliding beams of protons and antiprotons (the energy of each beam was 270 GeV) were carried out at the European Centre of Nuclear Research (CERN, Geneva). In these experiments, the particles with a mass exceeding that of the proton almost by a factor of 100 were obtained. Experiments with a fixed target and the same “useful energy” would require an accelerator designed for an energy of 155 TeV!

It would be wrong, however, to assume that only colliding-beam accelerators with storage rings should be designed. Fixed-target accelerators, which are inferior to storage-ring accelerators as regards the available energy, have a number of important advantages. Above all, it concerns the possibility to

**Table 12.** Large-Scale Accelerators**A. Fixed-Target Accelerators**

| Accelerator                                                                  | Year of commissioning | Particles being accelerated | Maximum energy (GeV) | Available energy (GeV) | Remark                                                                                   |
|------------------------------------------------------------------------------|-----------------------|-----------------------------|----------------------|------------------------|------------------------------------------------------------------------------------------|
| SLAG<br>Stanford, USA                                                        | 1961                  | $e^-$                       | 24                   | 5.8                    |                                                                                          |
| PS (CERN)<br>Geneva, Switzerland                                             | 1959                  | $p$                         | 28                   | 5.5                    |                                                                                          |
| AGS (BNL), USA                                                               | 1960                  | $p$                         | 33                   | 6.2                    |                                                                                          |
| Proton synchrotron (Inst. of High-Energy Physics),<br>Protvino, USSR         | 1967                  | $p$                         | 76                   | 10                     |                                                                                          |
| Proton synchrotron<br>Tevatron (FNAL)<br>Batavia, USA                        | 1972<br>1983          | $p$<br>$p$                  | 500<br>800           | 29<br>37               | 500 GeV accelerator is converted into Tevatron (the energy will be elevated to 1000 GeV) |
| SPS (CERN) Geneva, Switzerland                                               | 1976                  | $p$                         | 450                  | 27                     |                                                                                          |
| SRA (storage-ring accelerator) (Inst. of High-Energy Physics) Protvino, USSR | under construction    | $p$                         | 3000                 | 73                     |                                                                                          |

**B. Colliding-Beam Accelerators with Storage Rings**

| Accelerator                                          | Year of commissioning | Colliding beams | Energy of each beam (GeV) | Available energy (GeV) | Remark                             |
|------------------------------------------------------|-----------------------|-----------------|---------------------------|------------------------|------------------------------------|
| CEPB<br>(Inst. Nuclear Physics)<br>Novosibirsk, USSR | 1978                  | $e^+ e^-$       | 7                         | 14                     |                                    |
| CESR<br>Cornell Univ. USA                            | 1978                  | $e^+ e^-$       | 5.5                       | 11                     |                                    |
| PETRA (DESY)<br>Hamburg, FRG                         | 1978<br>1983          | $e^+ e^-$       | 19<br>23                  | 38<br>46               | Energy of beams has been increased |
| PEP (SLAC)<br>Stanford, USA                          | 1980                  | $e^+ e^-$       | 15                        | 30                     |                                    |
| ISR (CERN)<br>Geneva, Switzerland                    | 1971                  | $pp$            | 31                        | 62                     | Dismantled in 1984                 |



Table 12. (cont.)

| Accelerator                                          | Year of commissioning | Colliding beams | Energy of each beam (GeV) | Available energy (GeV) | Remark                                |
|------------------------------------------------------|-----------------------|-----------------|---------------------------|------------------------|---------------------------------------|
| SPS-collider<br>(CERN) Geneva, Switzerland           | 1981                  | $\bar{p}p$      | 270                       | 540                    | Energy of beams was increased in 1985 |
|                                                      | 1985                  |                 | 450                       | 900                    |                                       |
| Tevatron-collider (FNAL)<br>Batavia, USA             | under construction    | $\bar{p}p$      | 1000                      | 2000                   |                                       |
| SLK (SLAC)<br>Stanford, USA                          | ditto                 | $e^+e^-$        | 50                        | 100                    |                                       |
| LEP (CERN)<br>Geneva, Switzerland                    | ditto                 | $e^+e^-$        | 50                        | 100                    | First stage                           |
| HERA (DESY)<br>Hamburg, FRG                          | ditto                 | $e^-p$          | $e^-$ 30<br>$p$ 820       | 314                    |                                       |
| SRA (Inst. of High-Energy<br>Physics) Protvino, USSR | ditto                 | $pp$            | 3000                      | 6000                   |                                       |

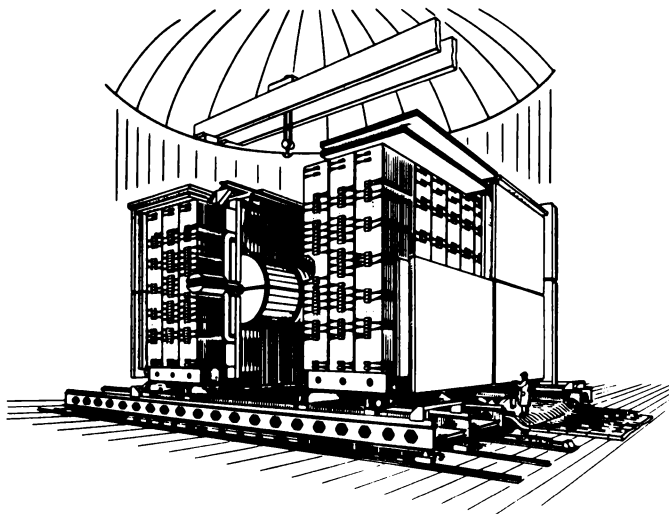


Fig. 422.

General view of the UA-1 experimental device intended for investigating  $\bar{p}p$ -collisions of the largest storage-ring accelerator operating on colliding proton and antiproton beams (SPS-collider, CERN, see Table 12). The UA-1 device is a huge magnetic spectrometer for measuring momenta of secondary particles formed as a result of  $\bar{p}p$ -collisions. The particles are registered in a gas-discharge chamber (it is seen in the middle of the device). The chamber is a system of a large number of gas-discharge detectors of particles resembling proportional counters. The detectors are designed for determining the trajectories of particles. The device also includes a large number of scintillation counters.

carry out experiments with various beams of unstable and neutral particles<sup>2</sup>, which is impossible for colliding-beam accelerators. Besides, fixed-target accelerators make it possible to investigate rare events since a considerably larger number of collisions can be obtained in them. For this reason, the researches with "ordinary" accelerators and with colliding-beam accelerators supplement each other, and taken together, provide a valuable information for the elementary-particle physics. Table 12 contains the main parameters of the largest accelerators operating and being designed in the world.

Besides large bubble chambers (see Sec. 25.6), experiments on modern ac-

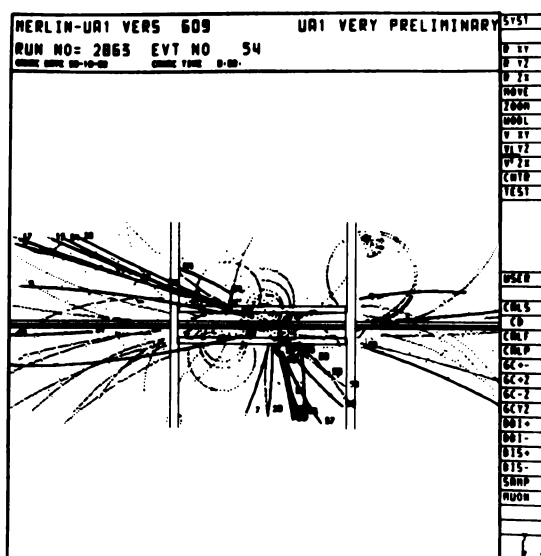


Fig. 423.

The photograph from the display of a computer operating together with the UA-1 device (see Fig. 422). It registers a  $\bar{p}p$ -collision at an energy of 270 GeV ( $\bar{p}$ ) + 270 GeV ( $p$ ). The information from all the detectors of the device processed by the computer makes it possible to determine the trajectories of the particles and to obtain the entire pattern of the interaction, which resembles photographs obtained in bubble chambers. The momenta of the particles were measured from the curvature of the tracks in the magnetic field. The photograph reflects the complexity of interactions occurring at such a high energy, in which a large number of secondary particles is formed.

<sup>2</sup> In experiments of fixed-target accelerators, the beams of secondary charged particles ( $\pi$ -mesons, protons, muons and other particles) with definite momenta are formed. The use is made of the deflection of charged particles in magnetic fields (which is inversely proportional to their momenta). Using magnetic fields with a complex configuration, it is possible to focus particle beams (in the same way as light beams are focussed by optical lenses). Neutral beams are separated by collimators and purified from charged particles with the help of magnetic fields.

celerators require large-scale and very complicated set-ups which can be compared with accelerators themselves as regards the size and complexity (Fig. 422). These devices include large magnetic spectrometers, thousands of fast scintillation counters and tens of thousands of gas-discharge detectors resembling the proportional counters mentioned in Sec. 23.3. These and other devices constituting experimental set-ups make it possible to determine the trajectories of particles, to measure their energy, momentum, velocity and ionization, to identify the particles and to investigate the interaction characteristics in detail. Such set-ups necessarily comprise a few computers required to process the obtained information, to adjust the numerous elements of the set-ups, to monitor their operation and to obtain the primary physical results allowing the researchers to control an experiment on the whole. Large body of information gained as a result of measurements is subjected to preliminary selection and then recorded on magnetic tapes. Later, the data are processed on large-scale high-speed computers. Figure 423 presents a photograph made from a computer display, which shows the form of an event recorded in the UA-1 device (see Fig. 422). This is an example of complex processes one has to deal with in a modern physical experiment.

## 26.2. Hadrons and Quarks

Investigations carried out on large-scale accelerators strongly contributed to the knowledge of elementary particles. Above all, this refers to the largest family of particles, *hadrons*, i.e. the particles participating in strong interactions. At present, a few hundreds of such particles are known, including *baryons* (the particles with a baryon charge  $B = +1$ ), *antibaryons* ( $B = -1$ ) and *mesons* whose baryon charge is zero. Most of these particles decay into other hadrons as a result of strong interactions. They have short lifetimes typical of nuclear processes ( $\sim 10^{-23}$  s, see Sec. 25.5). Such short time intervals cannot be measured directly. However, there are hadrons having a lifetime of about  $10^{-8}$ - $10^{-13}$  s. The decays of such long-lived (on the nuclear scale) particles are governed by weak interactions.

When only a small number of elementary particles was known, they were treated as “bricks” of the universe from which all the variety of atoms was built. At present, the number of elementary particles exceeds the number of chemical elements, and the very concept of elementary particle has clearly lost its meaning as regards hadrons.

There is no complete theory in the elementary-particle physics that would allow us to explain all main regularities and to attain the same level of understanding as that existing in classical mechanics or electrodynamics. In such a situation, the attempts of phenomenological analysis and classification of physical phenomena on the basis of certain conservation laws acquire a special significance. These laws help determine whether or not a given process may occur in nature.

Let us recall, for example, the law of conservation of baryon charge mentioned in the previous chapter. According to this law, the difference between the number of baryons and antibaryons is conserved in all processes. In order to express this law analytically, we ascribed a baryon charge  $B = +1$  to baryons and  $B = -1$  to antibaryons, the baryon charge of other particles being zero. Then the conservation of the number of baryons just means the conservation of the baryon charge.

In order to find out whether or not a given reaction is possible, we must first of all verify whether or not the electric and baryon charges are conserved in this reaction. Let us consider, for example, the process

$$p + p \rightarrow p + p + p + \bar{p}. \quad (26.2.1)$$

The initial particles have the total baryon charge  $\sum_{\text{in}} B = +1 + 1 = 2$ . For the particles in the final state, we have  $\sum_{\text{fin}} B = 1 + 1 + 1 - 1 = 2$ .

In other words, the baryon charge in the initial and final states is the same  $\left(\sum_{\text{in}} B = \sum_{\text{fin}} B\right)$ , and the reaction is possible. It can easily be verified that this reaction is also allowed by the electric charge conservation law (the electric charge of a proton is  $+1$  and of an antiproton,  $-1$ ). However, the reaction

$$p + n \rightarrow p + p + \bar{p} \quad (26.2.2)$$

turns out to be forbidden because of the nonconservation of baryon charge  $\left(\sum_{\text{in}} B = 2 \neq \sum_{\text{fin}} B = 1\right)$ , although the electric charge is conserved in it.

The other conservation laws will be considered below.

The establishment of the internal structure of elementary particles is one of the most important problems in modern physics. In order to solve this problem, it is very important to properly systematize the particles in the form of a table similar to the Periodic Table in some respects.

The first step in this direction was made when it was found that hadrons are grouped into small families having close properties.<sup>3</sup> Individual members of such families differ mainly in their electromagnetic properties, i.e. charges and magnetic moments. The examples of such families are the well-known nucleons (protons and neutrons) and  $\pi$ -mesons ( $\pi^+$ ,  $\pi^-$  and  $\pi^0$ ). However, the number of isotopic families is also very large (more than one hundred). These families in turn are combined into wider and more complex groups. The particles constituting such groups exhibit noticeable similarity, although they are not close relatives as the members of the same isotopic family. Such unions are based on a certain similarity or relation between the

<sup>3</sup> These families are known as *isotopic families* because of a certain analogy with the isotopes of elements, which have close properties.

basic parameters characterizing the particles. These parameters are known as *quantum numbers* of elementary particles.

The quantum numbers of hadrons are first of all their masses, electric charges, spins, magnetic moments, lifetimes and the values of baryon charge. The baryon and electric charges are not unique types of “charges” characterizing strongly interacting particles. It was established experimentally that in some reactions hadrons are generated in groups consisting of two or even more particles. Here we have a similarity with the formation of baryons and antibaryons which, as was mentioned above, are always born in pairs. The regularities associated with the formation of baryon-antibaryon pairs together with the information about the stability of nucleons revealed that baryons are characterized by a conserved quantum number called the baryon charge. However, the birth of the groups of new particles cannot be explained with the help of the laws of conservation of electric and baryon charges. Experiments showed that there are processes in which a proton is transformed into another baryon (so that the baryon charge is conserved), but it necessarily involves the formation of new types of mesons. This suggested a hypothesis that some hadrons have new specific quantum numbers, i.e. new “charges” which resemble to a certain extent the baryon charge and may have discrete positive, negative and zero values. These new “charges” were called *flavours*. These flavours were termed *strangeness*, *charm*, *beauty*, and so on.

Some of these names have a historical background. For example, in the fifties the first unusual particles were discovered, whose properties seemed enigmatic from the viewpoint of the existing concepts. Hence the name *strange particles*. When the unusual properties were explained by introducing a new quantum number, this new “charge” was called strangeness. On the whole, the abundance of exotic terms in elementary-particle physics (like quark, flavour, strangeness and charm) reflects the tendency of the physicists engaged in this field to use striking and memorable words which sound enigmatically and beautifully in all languages and at the same time reflect the fact that the nature of such objects is not clear so far and may conceal many unexpected things.

The general properties of some strongly interacting particles are represented in Table 13 which will be considered in greater detail somewhat later. This table, however, contains just a small part of the hadrons known at present, i.e. only comparatively long-lived particles that decay due to weak interactions (or under the action of electromagnetic forces). As was mentioned above most hadrons decay because of strong interactions, and their lifetimes range between  $10^{-22}$  and  $10^{-23}$  s. It is important that these short-lived hadrons do not differ in principle from long-lived particles.

In Table 13, we limit ourselves only to long-lived particles since otherwise the table would be as large as a booklet.

Table 13. Some Hadrons

| Name                        | Symbol      | Electric charge | Isotopic family      | Baryon charge | Stran | Charm | Beauty | Mass (MeV/c <sup>2</sup> ) | Lifetime (s)           | Anti-particle      | Quark composition                                                 |
|-----------------------------|-------------|-----------------|----------------------|---------------|-------|-------|--------|----------------------------|------------------------|--------------------|-------------------------------------------------------------------|
| "Ordinary Mesons particles" | Pi-plus     | $\pi^+$         | } $\pi$ -mesons      | 0             | 0     | 0     | 0      | 139.6                      | $2.60 \times 10^{-8}$  | $\pi^-$            | $\pi^+ = [u\bar{d}]$                                              |
|                             | Pi-zero     | $\pi^0$         |                      |               |       |       |        | 135.0                      | $0.80 \times 10^{-16}$ | $\pi^0$            | $\pi^0 = [\bar{u}u]$                                              |
|                             | Eta-meson   | $\eta$          | } $\eta$ -meson      | 0             | 0     | 0     | 0      | 549                        | $\sim 10^{-18}$        | $\eta$             | $\eta = [\bar{u}u]$<br>$\eta = [\bar{d}d]$<br>$\eta = [s\bar{s}]$ |
| Baryons                     | Proton      | $p$             | } $p, n$ -nucleons   | 1             | 0     | 0     | 0      | 938.3                      | $> 10^{32}$ years      | $(\bar{p})^-$      | $p = [uud]$                                                       |
|                             | Neutron     | $n$             |                      |               |       |       |        | 939.6                      | $\approx 900$          | $\bar{n}$          | $n = [udd]$                                                       |
| Mesons                      | K-plus      | $K^+$           | } $K$ -mesons        | 0             | +1    | 0     | 0      | 493.7                      | $1.24 \times 10^{-8}$  | $K^-$              | $K^+ = [u\bar{s}]$                                                |
|                             | K-zero      | $K^0$           |                      |               |       |       |        | 497.7                      | $0.89 \times 10^{-10}$ | $\bar{K}^0$        | $K^0 = [d\bar{s}]$                                                |
| Strange Baryons (hyperons)  | Lambda      | $\Lambda$       | } $\Lambda$ -hyperon | 1             | -1    | 0     | 0      | 1115.6                     | $2.6 \times 10^{-10}$  | $\bar{\Lambda}$    | $\bar{\Lambda} = uds$                                             |
|                             | Sigma-plus  | $\Sigma^+$      |                      |               |       |       |        | 1189.4                     | $0.80 \times 10^{-10}$ | $(\bar{\Sigma})^-$ | $\Sigma^+ = uus$                                                  |
|                             | Sigma-zero  | $\Sigma^0$      |                      |               |       |       |        | 1192.5                     | $6 \times 10^{-20}$    | $(\bar{\Sigma})^0$ | $\Sigma^0 = uds$                                                  |
|                             | Sigma-minus | $\Sigma^-$      |                      |               |       |       |        | 1197.3                     | $1.48 \times 10^{-10}$ | $(\bar{\Sigma})^+$ | $\Sigma^- = dds$                                                  |
|                             | Xi-zero     | $\Xi^0$         | } $\Xi$ -hyperons    | 1             | -2    | 0     | 0      | 1315                       | $2.9 \times 10^{-10}$  | $(\bar{\Xi})^0$    | $\Xi^0 = [u\bar{s}s]$                                             |
|                             |             | $\Xi^-$         |                      |               |       |       |        | 1321                       | $1.64 \times 10^{-10}$ | $(\bar{\Xi})^+$    | $\Xi^- = [d\bar{s}s]$                                             |
| Omega-minus particles       | Omega-minus | $\Omega^-$      | } $\Omega$ -hyperon  | 1             | -3    | 0     | 0      | 1672                       | $0.82 \times 10^{-10}$ | $(\bar{\Omega})^+$ | $\Omega^- = [sss]$                                                |

**Table 13. (cont.)**

| Name                               | Symbol        | Electric charge | Isotopic family                                                            | Baryon charge | Strangeness | Beauty | Mass (MeV/c <sup>2</sup> ) | Lifetime (s)                                                | Anti-particle         | Quark composition     |
|------------------------------------|---------------|-----------------|----------------------------------------------------------------------------|---------------|-------------|--------|----------------------------|-------------------------------------------------------------|-----------------------|-----------------------|
| Mesons                             | $D^+$         | +1              | $\left. \begin{array}{l} D\text{-mesons} \\ D^+, D^0 \end{array} \right\}$ | 0             | 0           | +1     | 0                          | $9 \times 10^{-13}$                                         | $D^-$                 | $D^+ = [cd]$          |
|                                    | $D^0$         | 0               |                                                                            | 0             | 0           | +1     | 0                          | $4 \times 10^{-13}$                                         | $\bar{D}^0$           | $D^0 = [c\bar{u}]$    |
|                                    | $F^+$         | +1              | $F\text{-meson}$                                                           | 0             | +1          | +1     | 0                          | $2 \times 10^{-13}$                                         | $F^-$                 | $F^+ = [cs]$          |
| Charmed Baryons (charmed hyperons) | $\Lambda_c^+$ | +1              | $\Lambda_c\text{-hyperons}$                                                | 1             | 0           | +1     | 0                          | $2 \times 10^{-13}$                                         | $(\bar{\Lambda}_c)^-$ | $\Lambda_c^+ = [udc]$ |
|                                    | $A^+$         | +1              | $A\text{-hyperons}$                                                        | 1             | -1          | +1     | 0                          | $2 \times 10^{-13}$                                         | $(\bar{A})^-$         | $A^+ = [usc]$         |
| Beautiful Mesons                   | $B^+$         | +1              | $\left. \begin{array}{l} B\text{-mesons} \\ B^+, B^0 \end{array} \right\}$ | 0             | 0           | 0      | +1                         | $\left. \begin{array}{l} 5270 \\ 5274 \end{array} \right\}$ | $B^-$                 | $B^+ = [ub]$          |
|                                    | $B^0$         | 0               |                                                                            | 0             | 0           | 0      | +1                         |                                                             | $\bar{B}^0$           | $B^0 = [d\bar{b}]$    |

*Remark:* The electric charges of particles are given in elementary charge units. Only a few charmed and beautiful particles are known so far, although the theory predicts the existence of a large number of such hadrons, both long-lived and short-lived.

A very large number of discovered hadrons and their grouping according to different classes and families make the elementary nature of such particles dubious. The grouping of hadrons into families and the nature and structure of these families, as well as other properties of hadron matter, have been explained in the most natural way by the *quark model of hadron structure*.

The fundamentals of this model can be formulated as follows.

1. Hadrons cannot be treated as elementary particles in the true sense of this word. They have a complex internal structure and, like atomic nuclei, are bound systems consisting of truly elementary, or basic particles. The basic structural elements constituting hadrons were called *quarks*.<sup>4</sup>

2. Hadron systematics (i.e. the study of the composition and properties of “related families” into which hadrons are grouped) makes it possible to state that all known baryons consist of three quarks ( $B = [q_1 q_2 q_3]$ ), antibaryons are composed of three antiquarks ( $\bar{B} = [\bar{q}_1 \bar{q}_2 \bar{q}_3]$ ), while all mesons consist of a quark and an antiquark ( $M = [q_1 \bar{q}_2]$ ). It turns out that quarks must have very unusual properties. Since the baryon charge is  $B = +1$  for baryons ( $B = -1$  for antibaryons), the quark structure of baryons implies that quarks have fractional baryon charge:  $B_q = 1/3$  and  $B_{\bar{q}} = -1/3$ . The electric charge of quarks must also be fractional (if the elementary charge is taken as unit charge):  $Q_q = +2/3$  or  $-1/3$  ( $Q_{\bar{q}} = -2/3$  or  $+1/3$ ). Only under these assumptions can the quantum numbers and properties of hadrons be explained.

3. There are at least six types of quarks each of which is a carrier of a new quantum number, viz. hadron flavour. These quarks are classified as follows:

|                                            |   |                                                                                               |
|--------------------------------------------|---|-----------------------------------------------------------------------------------------------|
| <i>u</i> -quark ( <i>up quark</i> )        | } | carriers of “isotopic flavours”                                                               |
| <i>d</i> -quark ( <i>down quark</i> )      |   | $I_z = +1/2$ for the <i>u</i> -quark and<br>$I_z = -1/2$ for the <i>d</i> -quark <sup>5</sup> |
| <i>s</i> -quark ( <i>strange quark</i> )   |   | carrier of the strangeness flavour $S = -1$                                                   |
| <i>c</i> -quark ( <i>charmed quark</i> )   |   | carrier of the charm flavour $C = +1$                                                         |
| <i>b</i> -quark ( <i>beautiful quark</i> ) |   | carrier of the beauty flavour $b = +1$                                                        |
| <i>t</i> -quark ( <i>true quark</i> )      |   | carrier of the truth flavour <sup>6</sup> $T = +1$ .                                          |

It should be emphasized that each quark carries only one flavour. All the remaining flavours are absent in it, i.e. the corresponding quantum numbers

<sup>4</sup> This term is due to the American physicist Murray Gell-Mann (b. 1929) who was the first to introduce the idea about the quark structure of hadrons.

<sup>5</sup> The question of isotopic flavours is more complicated, and we shall not consider their values here.

<sup>6</sup> The choice of the sign of different “charges” is always conditional. The quantity  $S = -1$  for the flavour of the strange quark was chosen arbitrarily and does not have any physical meaning. The *t*-quark is sometimes called the top quark.



are zero. Antiquarks differ from quarks in the opposite signs of all charges. For example, the  $s$ -quark is characterized by the electric charge  $Q_s = -1/3$ , baryon charge  $B_s = +1/3$  and strangeness  $S = -1$ , while all the other flavours are absent, i.e.  $I_z = 0$ ,  $C = 0$ ,  $b = 0$  and  $T = 0$ . For its antiquark, we have  $Q_s = +1/3$ ,  $B_s = -1/3$ ,  $S = +1$  and  $I_z = C = b = T = 0$ . The values of quantum numbers of quarks are given in Table 14.

4. Strong and electromagnetic interactions cannot change the individual properties of a quark, i.e. cannot change the values of the quark flavours. In other words, the laws of conservation of flavours are observed in these interactions (like the law of conservation of baryon charge). In the processes governed by strong and electromagnetic interactions, either quarks can be regrouped, or quark-antiquark pairs having definite flavours can be created (annihilated), or both events are possible.

5. Weak interactions play a unique role in nature since they change the individuality of quarks and may transform a quark having some flavour into a quark characterized by a different flavour. Thus, although flavours slightly resemble the baryon charge, there is an essential difference between them. The baryon charge is conserved in all processes known so far, while flavours exhibit much lower "stability" and are conserved only in strong and electromagnetic interactions.

The search for quarks having such distinct and peculiar properties in free state was undertaken in a large number of ingenious experiments. For example, one of the most sensitive experiments of this type was carried out on the Serpukhov accelerator soon after it had been put in operation. Another elegant experiment in which particles with fractional charges were sought in the surrounding medium was an improved version of the Millikan experiment on determining the elementary charge (see Sec. 22.4). It was carried out by physicists at the Moscow State University. However, quarks have not been observed in any of these and other numerous experiments.

At the same time, the analysis of the properties of hadrons convincingly proved that they indeed have a complex structure and consist of quarks. This was confirmed by experiments in which the spatial distribution of the electric charge and magnetic moment was investigated and which revealed the internal motion of quarks in hadrons. Moreover, the electric charges of quarks in hadrons were measured by using indirect methods, and it was found that these charges are indeed fractional, as was proposed earlier. A number of relations between the probabilities of formation or decay of strongly interacting particles and many other facts justify that the quark model is correct. By using this model, the existence of a number of new particles having quite definite properties was predicted, and these predictions were brilliantly confirmed in experiments. The rich body of experimental data evidence that quarks are a physical reality.

Table 14. Truly Elementary Particles

| Family  | 1st generation of basic particles                                    | 2nd generation of basic particles                                | 3rd generation of basic particles                              | Electric charge | Remark                                                                                                                                                                                                                                                                                                                                                                                         |
|---------|----------------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------------|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Quarks  | <i>u</i> -quark<br>$I_z = +1/2$ ,<br>$m \approx 350 \text{ MeV}/c^2$ | <i>c</i> -quark<br>$C = +1$ ,<br>$m \approx 1.5 \text{ GeV}/c^2$ | <i>t</i> -quark<br>$T = +1$ ,<br>$m > 23 \text{ GeV}/c^2$      | $Q_q = 2/3$     | For all quarks, the baryon charge $B = 1/3$ , the lepton charges being zero. Their spins are $1/2$ . Each type of quark has a corresponding antiquark differing from it in the signs of all charges and flavours. Quarks are kept within hadrons and are not observed in free state                                                                                                            |
|         | <i>d</i> -quark<br>$I_z = -1/2$ ,<br>$m \approx 350 \text{ MeV}/c^2$ | <i>s</i> -quark<br>$S = -1$ ,<br>$m \approx 5 \text{ GeV}/c^2$   | <i>b</i> -quark<br>$b = +1$ ,<br>$m \approx 5 \text{ GeV}/c^2$ | $Q_q = -1/3$    |                                                                                                                                                                                                                                                                                                                                                                                                |
|         | electron leptons<br>( <i>e</i> )<br>$l_e = 1, l_\mu = l_\tau = 0$    | muon leptons<br>( $\mu$ )<br>$l_\mu = 1, l_e = l_\tau = 0$       | tau-leptons<br>( $\tau$ )<br>$l_\tau = 1, l_e = l_\mu = 0$     |                 |                                                                                                                                                                                                                                                                                                                                                                                                |
|         | electrons ( $e^-$ )<br>$m = 0.511 \text{ MeV}/c^2$                   | muon ( $\mu^-$ )<br>$m = 105.7 \text{ MeV}/c^2$                  | $\bar{\tau}^-$ -leptons<br>$m = 1784 \text{ MeV}/c^2$          | $Q_l = -1$      |                                                                                                                                                                                                                                                                                                                                                                                                |
| Leptons | electron neutrino<br>( $\nu_e$ )<br>$m < 46 \text{ eV}/c^2$          | muon neutrino<br>( $\nu_\mu$ )<br>$m < 0.5 \text{ MeV}/c^2$      | tau-neutrino<br>( $\nu_\tau$ )<br>$m < 70 \text{ MeV}/c^2$     | $Q_l = 0$       | For all leptons, the baryon charge and quark flavours are zero. Leptons do not participate in strong interactions. Their spins are $1/2$ . Each type of lepton has its antilepton differing from the lepton in the signs of all charges. The value of the neutrino mass has not been determined so far, and we can only indicate the experimentally obtained upper limits for these quantities |

Table 14. (cont.)

| Family              | Basic properties                                                                                     | Electric charge | Remarks                                                                          |
|---------------------|------------------------------------------------------------------------------------------------------|-----------------|----------------------------------------------------------------------------------|
| Photons             | The rest mass is zero, the spin is unity                                                             | $Q = 0$         | Photons are quanta of the electromagnetic field                                  |
| Gluons              | The rest mass is zero, the spin is unity. Gluons exist in hadrons and are not observed in free state | $Q = 0$         | Gluons are quanta of the strong interaction field, keeping quarks within hadrons |
| Intermediate bosons | $W^\pm$ -bosons, $m = 81 \text{ GeV}/c^2$ , the spin is unity                                        | $Q_W = \pm 1$   | Intermediate bosons are quanta of the weak force field                           |
|                     | $Z^0$ -bosons, $m = 93 \text{ GeV}/c^2$ , the spin is unity                                          | $Q_Z = 0$       |                                                                                  |

*Remark.* For each quark, the value of the flavour carried by it is indicated (all the remaining flavours are zero). The flavours  $I_z = 1/2$  and  $I_z = -1/2$  for  $u$ - and  $d$ -quarks are known as their isotopic flavours. These flavours determine the grouping of particles into isotopic families having very close properties. The quarks are referred to as follows:  $u$ -quark is the up quark,  $d$ -quark is the down quark,  $s$ -quark is the strange quark,  $c$ -quark is the charmed quark,  $b$ -quark is the beautiful quark and  $t$ -quark is the true quark (this term has not been generally accepted).

Since quarks can exist only within hadrons, their masses have approximate values because we can speak about the mass of a component of a system only if the mass defect is small. The large mass of  $c$ -,  $b$ - and  $t$ -quarks determines large masses of charmed, beautiful and  $t$ -particles.

The existence of particles having  $t$ -quarks has not yet been confirmed completely. The  $\nu_\tau$ -neutrino has not been observed directly in experiments, and its existence has been established on the basis of indirect indications.

The spin of any particle expressed in special quantum units ( $\hbar/2\pi$ ) can be either integral or half-integral, which constitutes its remarkable property.

But why do they exist within hadrons and are not observed in free form? There is no unambiguous answer to this question at the moment. It has been established, however, that quarks are bound by a special type of force due to the exchange by gluons which are not observed in free state either. These forces “glue” quarks in hadrons and apparently do not allow quarks to leave hadrons during any collisions. Hadrons may “split” with the formation of many other hadrons. In other words, a large number of quark-antiquark pairs, which are then bound into compound particles, are generated in the process of collision. However, free quarks never escape from an initial hadron. The situation observed here resembles in some respects the experiments with permanent magnets: applying a tensile force to a magnet, we can break it and obtain new magnetic dipoles, but it is impossible to isolate magnetic poles.

The problem of quarks and gluons that cannot escape from a hadron has become known as *confinement* (imprisonment) and is one of the most fundamental problems of elementary-particle physics that awaits its solution.

### 26.3. Quark Structure of Hadrons

The basic concepts of the quark model formulated in the previous section make it possible to explain qualitatively all main properties of hadron phenomena. Let us apply these concepts to hadron systematics and determine which types of strongly interacting particles can exist according to the quark concepts.<sup>7</sup> Let us first consider "ordinary particles" consisting only of  $u$ - and  $d$ -quarks.

Let us begin with the lightest baryons, viz. protons and neutrons. The correct values of quantum numbers for these particles can be obtained by assuming that they have the following quark structure:  $p = [uud]$  and  $n = [udd]$ . Indeed, then the baryon charges of these particles are  $B_p = B_n = 3 \times (1/3) = 1$ , and the electric charges are  $Q_p = 2Q_u + Q_d = 2 \times (2/3) - 1/3 = 1$  and  $Q_n = Q_u + 2Q_d = 2/3 - 2 \times (1/3) = 0$ . Protons differ from neutrons in their electric charge and in the values of isotopic flavours. As was mentioned in this and the previous chapters, protons and neutrons form the isotopic family of nucleons, i.e. the particles having very close properties. In order to explain this similarity, we have to assume that strong interactions between the  $u$ - and  $d$ -quarks are similar and that the quark systems which differ from one another only in that  $u$ - and  $d$ -quarks change places have similar properties and form the isotopic family of particles. Individual members of such an isotopic family can be considered to be different charged states of the same particle (in the case under consideration, the proton and the neutron are different states of the nucleon). Such a model can be confirmed by analyzing the properties of mesons consisting of  $u$ - and  $d$ -quarks and the corresponding antiquarks. Here there can be systems with zero baryon charge  $[u\bar{d}]$ ,  $[d\bar{u}]$  and  $[u\bar{u}] \rightleftharpoons [d\bar{d}]$ .<sup>8</sup> They have electric charges  $+1$ ,  $-1$  and  $0$ . The lightest of such systems are the well-known  $\pi^+$ -,  $\pi^-$ - and  $\pi^0$ -mesons which also form the isotopic family of  $\pi$ -mesons. It is interesting to note that  $\pi^+$ - and  $\pi^-$ -mesons are a particle and an antiparticle (it follows from their quark structure). As a truly neutral particle (i.e. the particle whose all charges are zero), the  $\pi^0$ -meson is identical to its antiparticle. The fact that a particle and an antiparticle belong to the same isotopic family is the common property of all the mesons consisting of  $u$ - and  $d$ -quarks.

Not all systems having the same quark composition are close in physical properties and contained in the same isotopic family. In the same way as atoms may have the ground and excited states, a quark system may have "excited states" characterized by larger values of mass besides the ground state with the minimum mass. If such "excited levels" are located sufficiently high and a system can go over to lower levels by emitting  $\pi$ -mesons, these transitions are governed by strong interactions. An "excited state" has a lifetime typical of strong interactions ( $\sim 10^{-23}$  s). As was mentioned earlier, a large number of such "excited" baryons and mesons which are also grouped in their isotopic families has been discovered. The isotopic families of "excited" particles may have a different structure. For example, among baryons there are groups of four particles  $\Delta^{++} = [uuu]$ ,  $\Delta^+ = [uud]$ ,  $\Delta^0 = [udd]$  and  $\Delta^- = [ddd]$ . They are called *Δ-isobars*. Among mesons, these are "families" consisting of only one particle.

Let us now consider hadrons which include, in addition to  $u$ - and  $d$ -quarks, the quarks having different flavours. They are known as strange particles, charmed particles, and so on. Let us demonstrate the basic features of the quark structure of such hadrons for the strange  $s$ -quarks that have been investigated to a much larger extent than charmed and beautiful particles.

The particles containing  $s$ -quarks have a nonzero strangeness. If there is only one strange quark in the composition of a strange baryon (or a *hyperon*), the hyperon strangeness  $S = -1$ .

<sup>7</sup> This section can be regarded as an explanation of Table 13.

<sup>8</sup> Since the neutral systems  $u\bar{u}$  and  $d\bar{d}$  may be transformed into each other (this process can be treated as the annihilation of the  $u\bar{u}$ -pair and the formation of the  $d\bar{d}$ -pair, or vice versa), the  $[u\bar{u}] \rightleftharpoons [d\bar{d}]$ -systems form a neutral compound particle which some part of its lifetime is in the  $[u\bar{u}]$ -state and the other part, in the  $[d\bar{d}]$ -state (i.e.  $\pi^0$ -meson).

Such a particle is a  $\Lambda$ -hyperon or the whole isotopic family consisting of three  $\Sigma$ -hyperons:  $\Lambda^0 = [uds]$  is a  $\Lambda$ -hyperon ( $Q_\Lambda = 0$ ,  $B_\Lambda = 1$ ), and  $\Sigma^+ = [uus]$ ,  $\Sigma^0 = [uds]$  and  $\Sigma^- = [dds]$  are  $\Sigma$ -hyperons differing from one another in charge and in isotopic flavour. Hyperons having two strange quarks are characterized by the strangeness  $S = -2$  and form the isotopic family consisting of two particles known as  $\Xi$ -hyperons. Their quark structure is  $\Xi^0 = [uss]$  and  $\Xi^- = [dss]$ , from which we can easily obtain the values of their quantum numbers. There is also the  $\Omega^-$ -hyperon having the strangeness  $S = -3$  and consisting only of three  $s$ -quarks:  $\Omega^- = [sss]$ . Since the  $\Omega^-$ -hyperon does not contain  $u$ - or  $d$ -quarks it does not have any “close relatives”, and the corresponding isotopic family consists of only one particle.

Let us now consider strange mesons. In this class there must be particles of the types  $[s\bar{u}]$  and  $[s\bar{d}]$ . The lightest of them (“ground states”) are known as  $K$ -mesons and form the isotopic family consisting of two particles:  $K^+ = [s\bar{u}]$  and  $K^0 = [s\bar{d}]$ . Their strangeness is  $S = +1$ . The quark structure of these mesons implies that their antiparticles form another isotopic family of two particles with  $S = -1$ . These are *anti-K-mesons*:  $K^- = [\bar{s}u]$  and  $\bar{K}^0 = [\bar{s}d]$ .

Besides the lightest strange baryons and  $K$ -mesons, there is a large number of their “excited states”, viz. heavier short-lived strange particles. The situation here is the same as with “ordinary” particles consisting of the  $u$ - and  $d$ -quarks.

Recently, other types of hadrons having new quantum numbers similar to strangeness, i.e. charmed and beautiful baryons and mesons, have been found. Besides the  $u$ -,  $d$ - and  $s$ -quarks, these particles contain the charmed  $c$ -quarks and the beautiful  $b$ -quarks. The quark model predicts a large variety of such particles which are formed by all possible combinations of three quarks (baryons) or of a quark and an antiquark (mesons). Just a few charmed and beautiful hadrons have been discovered so far. The information about the already found particles and their quark structure is contained in Table 13.

As a rule, strange particles have somewhat larger masses than “ordinary” particles. The mass of charmed particles is much larger than that of strange particles (their masses are  $2\text{--}3 \text{ GeV}/c^2$ ), while beautiful particles are characterized by still larger masses ( $> 5 \text{ GeV}/c^2$ ). Such a difference in the masses of these particles is attributed to the difference in the masses of quarks constituting them (see Table 14). Recently, the results have been obtained in favour of the existence of one more class of hadrons with very large masses of  $30\text{--}40 \text{ GeV}/c^2$ . These hadrons apparently contain heavy  $t$ -quarks.

It was shown earlier that the similarity in the properties of the  $u$ - and  $d$ -quarks has led to the existence of isotopic families of hadrons having very close properties. The  $s$ -quark differs from the  $u$ - and  $d$ -quarks but not very strongly. This allows us to explain why the individual isotopic families of hadrons comprising the  $u$ -,  $d$ - and  $s$ -quarks are combined into the related groups of particles mentioned in the previous section. A more detailed analysis based on the quark model makes it possible to determine the composition and some properties of such groups, which turn out to be in excellent agreement with experiment. Thus, the quark model allows us to explain the main features of the hadron systematics.

## 26.4. Quark Model and Formation and Decay of Hadrons

Using the rules of the quark model (see Sec. 26.2) and the data compiled in Tables 13 and 14 (see Secs. 26.2 and 26.3), we shall briefly consider the description of various hadron reactions in terms of quarks. For example, let us consider again the formation of  $\pi$ -mesons in the nucleon-nucleon interactions  $n + p \rightarrow n + n + \pi^+$  (see Sec. 25.3). In terms of the quark model, this reaction can be written as follows:

$$\underbrace{[udd]}_n + \underbrace{[uud]}_p \rightarrow \underbrace{[udd]}_n + \underbrace{[udd]}_n + \underbrace{[u\bar{d}]}_{\pi^+}. \quad (26.4.1)$$

It can be seen that the formation of a  $\pi$ -meson has been reduced to the formation of the quark-antiquark pair  $d\bar{d}$  and to the regrouping of the quarks. Such a reaction satisfies the rules of the quark model (see Sec. 26.2, item 4) and hence is possible. Another example is the formation of baryons and antibaryons  $\pi^- + p \rightarrow \pi^- + p + p + \bar{p}$  (such an interaction is recorded on the photograph made in the bubble chamber and represented in Fig. 416). In the quark model, we have the following process:

$$\underbrace{[d\bar{u}]}_{\pi^-} + \underbrace{[uud]}_p \rightarrow \underbrace{[d\bar{u}]}_{\pi^-} + \underbrace{[uud]}_p + \underbrace{[uud]}_p + \underbrace{[\bar{u}\bar{u}\bar{d}]}_{\bar{p}}, \quad (26.4.2)$$

i.e. the formation of two  $u\bar{u}$ -pairs and a  $d\bar{d}$ -pair, which are then grouped into a proton and an antiproton.

Let us now consider the formation of strange particles. Here too only such processes are allowed that lead to the formation (annihilation) of quark-antiquark pairs having a certain flavour and to the quark regrouping. For example, the reaction

$$\underbrace{[\bar{u}d]}_{\pi^+} + \underbrace{[uud]}_p \rightarrow \underbrace{[uus]}_{\Sigma^+} + \underbrace{[\bar{s}u]}_{K^+} \quad (26.4.3)$$

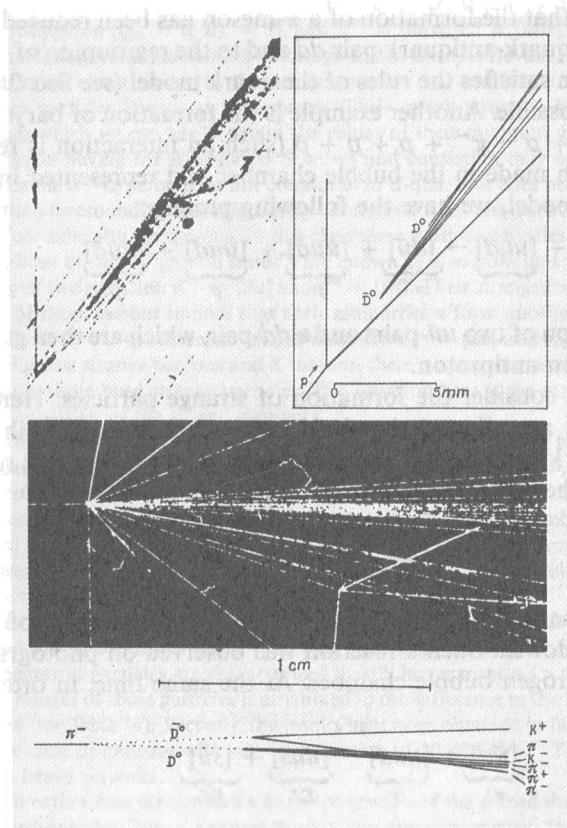
is reduced to the annihilation of a  $d\bar{d}$ -pair and to the creation of an  $s\bar{s}$ -pair, and hence is allowed. Such a reaction was observed on photographs taken in a liquid-hydrogen bubble chamber. At the same time, in order that the reaction

$$\underbrace{[\bar{u}d]}_{\pi^+} + \underbrace{[uud]}_p \rightarrow \underbrace{[uus]}_{\Sigma^+} + \underbrace{[s\bar{u}]}_{K^-} \quad (26.4.4)$$

take place, the quarks must change: two  $s$ -quarks should be transformed into two  $u$ -quarks. According to the basic concepts of the quark model, such processes are not allowed, at any rate for the strong and electromagnetic interactions in which flavours are conserved. Indeed, reaction (26.4.4) has never been observed in any experiment.

We can also consider some other reactions like those leading to the simultaneous creation of the whole groups of strange particles. Such an analysis is very simple. The examples considered above show that it is essentially reduced to a certain "play with blocks" where the blocks are quarks with certain flavours. The "rules of play" here are formulated as postulates of the quark model (see Sec. 26.2).

Recently, the events involving the simultaneous formation of charmed and beautiful hadrons have been observed in many strong and electromagnetic processes at high energies. The decays of such particles are due to weak interactions and are characterized by lifetimes of the order of  $10^{-12}$ - $10^{-13}$  s. In such processes, the particles have time to cover only very short distances

**Fig. 424.**

Formation and decay of charmed  $D^0$ - and  $\bar{D}^0$ -mesons. The figure represents two photographs of the formation of pairs of charmed particles in the reaction  $\pi^- + p \rightarrow D^0 + \bar{D}^0 + (\text{other particles})$  in bubble chambers with large spatial resolutions. The lifetime of the  $D^0$ -mesons is of the order of  $\sim 10^{-12}$  s, during which they traverse distances of a few millimetres. The decays  $D^0 \rightarrow K^- + \pi^+ + \pi^+ + \pi^-$  and  $\bar{D}^0 \rightarrow K^+ + \pi^-$  have been registered. The schematic diagrams of these events are given at the right top and at the bottom of the figure.

of the order of a few millimetres even on account of the relativistic increase in their lifetimes (see Sec. 25.7). Such particles can be registered only by the detectors capable to reliably measure such short distances. Figure 424 shows the photographs of the events occurring during the  $\pi^- p$ -interactions and obtained in special small bubble chambers in which it is possible to take very accurate photographs with a good spatial resolution. On these photographs, we can see the formation of a pair of charmed particles, viz.  $D^0$ - and

$\bar{D}^0$ -meson, accompanied by other hadrons (mainly  $\pi$ -mesons)

$$\pi^- + p \rightarrow D^0 + \bar{D}^0 + (\text{other particles}). \quad (26.4.5)$$

From the point of view of the quark model, the reaction is reduced to the formation of a certain number of quark-antiquark pairs (including a pair of the  $c\bar{c}$ -quarks) which are then grouped into the  $D^0$  and  $\bar{D}^0$ -mesons and additional hadrons. Generally, it should be noted that at high energies, collisions may result in the formation of a very large number of the quark-antiquark pairs which are then manifested in the multiple formation of hadrons. One of such events occurring in colliding beams of protons and antiprotons with a very high admissible energy of 540 GeV was shown in Fig. 423.

Quark flavours are not strictly conserved quantum numbers and can change in weak interactions. Weak decays of hadrons are therefore due to transformations of quarks having certain flavours into quarks with different flavours (see Sec. 26.2, item 5). For example, the decay of strange  $\Lambda$ -hyperons through the channel  $\Lambda \rightarrow p + \pi^-$  (see the photograph in Fig. 419) is observed. In terms of the quark model, this process can be described by two stages. Weak interactions lead to the transformation  $s \rightarrow u$  in which an  $s$ -quark is converted into a  $u$ -quark having another flavour. A “weak” formation of the quark-antiquark pair  $d\bar{u}$  with different quark flavours also takes place.<sup>9</sup> Thus, the first stage of the hyperon decay is reduced to the “weak” transformation  $s \rightarrow u + d + \bar{u}$ . On the second stage, quarks are regrouped due to strong interactions leading to the formation of two hadrons, viz. a proton and a  $\pi^-$ -meson:

$$[uds] \xrightarrow[\text{transformation } s \rightarrow u\bar{d}\bar{u}]{\substack{\text{1st stage} \\ \text{“weak”}}} \boxed{udud\bar{u}} \xrightarrow[\text{interactions}]{\substack{\text{2nd stage} \\ \text{strong}}} \underbrace{[uud]}_p + \underbrace{[\bar{u}d]}_{\pi^-}. \quad (26.4.6)$$

This example shows that strong interactions also play a definite role in weak decays of hadrons resulting in the formation of strongly interacting particles in the final state. However, such weak decays are based on a weak process causing a transformation of the initial quarks.

Concluding the section, let us consider the  $\beta$ -decay of neutrons  $n \rightarrow p + e^- + \bar{\nu}_e$ , which was mentioned more than once in this book (see Secs. 25.1 and 25.4). Weak interactions cause here the transformation of a  $d$ -quark into a  $u$ -quark accompanied by the formation of the leptons  $e^-$  and

<sup>9</sup> Let us emphasize once again that this process differs in principle from strong interactions in which only quark-antiquark pairs with opposite signs of the same flavour can be created.



$\nu_e$ :

$$d \rightarrow u + e^- + \bar{\nu}_e$$

$$\underbrace{[udd]}_n \rightarrow \underbrace{[uud]}_p + e^- + \bar{\nu}_e. \quad (26.4.7)$$

This example shows that weak forces change the individuality of not only quarks but also leptons, forming pairs of leptons of various types (leptons will be considered in greater detail in the next section).

## 26.5. Leptons. Intermediate Bosons.

### The Unity of All Interactions

The rapid development of elementary-particle physics during recent years has considerably changed our ideas not only about hadrons, but also about leptons, i.e. the particles exhibiting only weak and electromagnetic (for charged leptons) interactions. Besides two pairs of leptons known earlier (electrons and electron neutrinos and muons and muon neutrinos, see Secs. 25.2, 25.5 and 25.6), one more charged lepton referred to as *tau-lepton* ( $\tau^-$ ) was discovered. Apparently, one more type of neutrino, viz. the tau-neutrino ( $\nu_\tau$ ) must exist together with the  $\tau$ -lepton. True, the latter has not yet been observed in direct experiments. Tau-neutrinos may appear, for example, during the decay of tau-leptons or be emitted together with tau-leptons during the decays of heavier particles.

Each lepton has a corresponding antiparticle, viz. an antilepton. Numerous experiments revealed that up to distances of the order of  $10^{-16}$  cm, leptons and antileptons behave as elementary "point-like" objects. It is believed now that it is leptons that together with quarks are truly elementary, or basic, particles (see Table 14).

All processes of lepton formation or decay (some of which were considered earlier, see Sec. 25.4) can be explained by assuming that leptons also have definite conserved quantum numbers known as lepton charges and resembling the baryon charge.

Three types of such lepton charges are known at present: electron ( $l_e$ ), muon ( $l_\mu$ ) and tau-lepton ( $l_\tau$ ) charges.

(1) The electron lepton charge  $l_e = +1$  for electrons  $e^-$  and electron neutrinos  $\nu_e$ . For their antiparticles ( $e^+$  and  $\bar{\nu}_e$ ),  $l_e = -1$ , while for other particles  $l_e = 0$ .

(2) The muon lepton charge  $l_\mu = +1$  for muons  $\mu^-$  and muon neutrinos  $\nu_\mu$ . For the corresponding antiparticles ( $\mu^+$  and  $\bar{\nu}_\mu$ ),  $l_\mu = -1$ , while for the remaining particles  $l_\mu = 0$ .

(3) The tau-lepton charge  $l_\tau = +1$  for the tau-lepton  $\tau^-$  and the tau-neutrino  $\nu_\tau$ . For anti-tau-leptons ( $\tau^+$  and  $\bar{\nu}_\tau$ ),  $l_\tau = -1$ , while for all other particles,  $l_\tau = 0$ .

In all the processes investigated so far, all three types of lepton charge are conserved. By way of an exercise, we suggest that the reader should prove by using the idea of conserved lepton charges that decays (25.4.1), (25.4.2) and reactions (25.4.3) and (25.4.4) may occur in nature, while the processes  $\bar{\nu}_e + n \rightarrow p + e^-$ ,  $\nu_\mu + n \rightarrow p + e^-$ ,  $\mu^- \rightarrow e^- + \gamma$  and  $\tau \rightarrow \mu^+ + e^+ + e^-$  are forbidden. Indeed, these and other transformations in which the laws of conservation of lepton charges are violated have never been observed in any of the numerous experiments in the quest of new particles. Leptons have neither baryon charge nor quark flavour, i.e. the corresponding quantum numbers are zero for them. This is due to the fact that leptons do not participate in strong interactions.

Table 14 contains the particles which are assumed truly elementary at present. Hadrons are not included in this table since their composite internal structure has been established quite reliably and it has been proved that quarks “glued together” by the gluon exchange are just the structural elements constituting hadrons. However, this table should be supplemented with other elementary particles. These are first of all photons, viz. the electromagnetic field quanta, which execute electromagnetic interactions between charged particles. Gluons are also included in this table since they execute interactions between quarks and together with them are condemned to “life imprisonment” within hadrons.

Weak interactions also play a very important role in elementary-particle physics. It was mentioned earlier that it is the only interaction in nature that can change the individuality of basic particles (leptons and quarks) and cause their mutual transformations obeying, however, the laws of conservation of the lepton and baryon charges). The mechanism of weak forces has attracted the attention of researchers for a long time. A hypothesis was put forth according to which these forces are due to the exchange of special type of quanta of the weak interaction force field, which were called *intermediate bosons*. Unlike gluons, intermediate bosons, as well as photons, must exist in free state. The theory has made it possible to predict the existence of three such intermediate bosons:  $W^\pm$ - and  $Z^0$ -particles. Finally, in 1982-83 intermediate bosons were discovered, and this discovery was a sensation.

Intermediate bosons were registered in complex experiments on storage-ring accelerator with colliding beams of protons and antiprotons having an energy of 270 GeV each (later this energy was increased to 450 GeV). This was the highest energy obtained artificially. The general view of one of the two giant machines on which this remarkable discovery was made is shown in Fig. 422. Figure 425 represents a photograph taken from the display of a computer on which the formation and decay of intermediate  $W$ -bosons was recorded.

The masses of intermediate bosons were found to exceed almost a

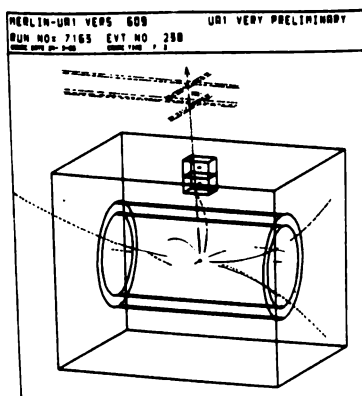


Fig. 425.

Formation and decay of intermediate bosons. The photograph is taken from the display of the computer on which the events registered in the UA-1 device have been processed (see Fig. 422). The beams of protons and antiprotons are directed along the axis of the cylindrical gas-discharge chamber of the device, which is shown schematically on the display. The event of the  $\bar{p}p$ -interaction is registered, in which a heavy intermediate boson  $W$  is formed. The event  $\bar{p} + p \rightarrow W + (\text{other particles})$  has been recorded. The decay  $W \rightarrow \mu + \nu_\mu$  is observed (the meson has a nearly transverse track with a large momentum). The neutrino flies in the opposite direction. It cannot be observed directly, but is identified from the kinematics of the event since  $\nu_\mu$  carries away a large momentum.

hundred times the masses of nucleons (see Table 14). These are the heaviest particles created in the laboratory conditions.

The discovery of intermediate bosons has crowned a very important cycle of investigations which have indicated that in spite of their apparent difference, weak and electromagnetic forces are closely related and are essentially the manifestations of the same interaction which is called *electroweak interaction*. Serious attempts are being made now to establish a relation between the electroweak and strong interaction and to try and explain the unique nature of all four types of forces existing in nature, i.e. strong, electromagnetic, weak and gravitational forces.

The idea about the "grand unification" of strong, electromagnetic and weak interactions is in contradiction with the division of basic particles into strongly interacting quarks, and leptons which do not take part in such interactions. Some common features of quarks and leptons possibly follow from their division into the groups having a similar structure. Table 14 shows that we can speak about three such groups, or generations, of basic particles: light  $u$ - and  $d$ -quarks and light leptons ( $e$  and  $\nu_e$ ) form the first such generation, heavier  $c$ - and  $s$ -quarks together with muons and muon neutrinos constitute the second generation, while the heaviest  $t$ - and  $b$ -quarks and leptons ( $\tau$  and  $\nu_\tau$ ) are included into the third generation. In all probability, there must

be processes in which quarks are transformed into leptons, while various types of leptons ( $e$ ,  $\mu$  and  $\tau$ ) also experience mutual transformations. The search for phenomena in which the conservation of baryon and lepton charges is violated (although with a very low probability) is of utmost importance for modern science. For example, the search for the decays of protons into lighter particles ( $p \rightarrow e^+ + \gamma$ ,  $p \rightarrow e^+ + \pi^0$ ,  $p \rightarrow \mu^+ + \pi^0$ ) is on in many laboratories of the world. Considerable amounts of energy should be liberated in such decays because of the large mass of the proton.

These experiments are carried out on complex devices with large "sensitive volumes" of a substance. The term "sensitive volume" means that if a nucleon decays into light particles in this volume, such a decay will be registered. The sensitive volumes of available and planned devices contain  $10^{31}$ - $10^{33}$  nucleons, and the exposure time is measured in years. In order to protect such devices from cosmic radiation, they are mounted in underground laboratories at a large depth. The decay of a proton has not yet been reliably registered so far. A few of detected events (so-called candidates to proton decays) can be due to background processes. It was established in these experiments that even if the proton is not absolutely stable, it has a huge lifetime  $\tau_p > 10^{31}$ - $10^{32}$  years. This means, for one, that none of the protons constituting a human body decays during the entire life of the person with a very high probability. The scale of the proton lifetime appears enormously large even as compared to the age of the Universe ( $\sim 10^{10}$  years).

# Answers and Solutions

## Part I

### Oscillations and Waves

**I.1.**  $g = 9.87 \text{ m/s}^2$ . The maximum value of  $g$  on the Earth (at the poles) slightly exceeds  $9.83 \text{ m/s}^2$ .

**I.2.**  $l = 895 \text{ m}$ .

**I.3.** The free fall acceleration on the equator and at a pole is  $9.780$  and  $9.832 \text{ m/s}^2$  respectively. Since the lengths of the pendulums are equal, the ratio of the periods is

$$\frac{T_{\text{equator}}}{T_{\text{pole}}} = \sqrt{\frac{9.832}{9.780}} = 1.002.$$

Thus, during the time required for 1000 pendulum oscillations on the equator, the pendulum at the pole will complete 1002 oscillations.

**I.4.**  $l = 99.5 \text{ m}$ .

**I.5.**  $l = 2h/\pi^2$ .

**I.6.** The ball rolls down the arc of the circle during a quarter of a period of the pendulum of length  $R$ , i.e. ( $h$  is small as compared to  $R$ ) over the time  $t_1 = (\pi/2)\sqrt{R/g}$ . The time required for the ball to roll down the chord is  $t_2 = \sqrt{2s/(g \sin \alpha)}$ , where  $s$  is the length of the chord and  $\alpha$  is its slope to the horizontal plane:  $\sin \alpha = h/s$ . Considering that  $s^2 = 2hR$ , we obtain  $t_2 = 2\sqrt{R/g}$ . Thus,  $t_2$  (as well as  $t_1$ ) does not depend on  $h$ . The velocities of the balls at the lower point will be equal:  $v = \sqrt{2gh}$ .

**I.7.** As the water flows out, the centre of gravity is lowered, and the length of the pendulum, and hence its period increase.

**I.8.** The period of a simple pendulum is  $T = 2\pi\sqrt{m/k}$ , where  $k$  is the rigidity of the spring, i.e. the proportionality factor between the tensile force  $F$  and spring elongation  $l$ :  $F = kl$ . If  $F$  is the force of gravity acting on the vibrating load, i.e.  $F = mg$ , we obtain

$$k = \frac{F}{l} = \frac{mg}{l}.$$

Substituting this expression for  $k$  into the formula for period  $T$ , we obtain

$$T = 2\pi \sqrt{\frac{l}{g}}.$$

For  $g = 9.81 \text{ m/s}^2$  and  $l = 0.002 \text{ m}$ , we have  $T = 0.09 \text{ s}$ .

**I.9.** Such impacts are equivalent to the action of the resultant of sinusoidal forces so that the lowest (fundamental) frequency is twice as high as the natural frequency of the pendulum. Consequently, neither the fundamental frequency of the force nor its higher harmonics can get in resonance.

**I.10.** The restoring force is  $mg \sin \varphi$ , where  $\varphi$  is the angle of deflection of the pendulum. For small angles this force is proportional to the angle since  $\sin \varphi \approx \varphi$  for small angles. If the force were proportional to angle for all angles (i.e. were equal to  $mg\varphi$ ), the isochronism would be conserved for all amplitudes. However, as the angle increases, the force increases at a slower rate than the angle ( $\sin \varphi < \varphi$ ). Therefore, with increasing amplitude the period will increase.

I.11. Systoles.

I.12. If we imagine that a square with vertical and horizontal sides is drawn on the screen, the spot will move (a) along a diagonal of the square, (b) along the other diagonal, (c) in a circle inscribed into the square.

I.13. 32 min and 1.28 s.

I.14. 7.5 times.

I.15. A light year is equal to  $9.5 \times 10^{12}$  km. A parsec is equal to  $3.1 \times 10^{13}$  km.

I.16. When a lightning is not far, the primary acoustic wave from the lightning itself is many times stronger than the echo arriving later as a result of reflection at various distant objects: clouds, woods, hills, and so on. When a lightning strikes at a large distance from an observer, the primary and reflected waves reach the observer with a smaller difference in strength.

I.17.  $\lambda = 89.6$  m.

I.18.  $T = 1/140$  s,  $\nu = 140$  Hz and  $\lambda = 2.4$  m.

I.19.  $l = 2.6$  mm at the beginning of the sound track and 1 mm at the end of the track.

I.20. A longitudinal acoustic wave propagating along the wire practically does not exhibit any dissipation of energy in the transverse directions, i.e. we have a directional propagation of sound (along the wire). The absorption of the wave in the wire material is also small.

I.21. As the distance between the sources increases, the number of alternating lines of maxima and minima increases. If the separation  $d$  of the sources lies between  $(n - 1/2)\lambda$  and  $(n + 1/2)\lambda$ , the number of the lines of minima will be  $2n$ . For  $d < \lambda/2$ , the vibrations will be enhanced everywhere since the path difference, which is never larger than  $d$ , is smaller than  $\lambda/2$  everywhere.

I.22. The pitch rises due to an increase in the speed of sound in air. If we neglect the very weak effect of expansion of instruments themselves, the increase in the pitch is the same for brass and wood instruments.

I.23. About 28 cm.

I.24. One should measure the depth of the channel in the key. It is equal to a quarter of the wavelength in air for the required frequency.

I.25. If the fundamental frequency of the tube open at both ends is  $\nu$ , the overtones will be  $2\nu$ ,  $3\nu$ ,  $4\nu$ ,  $5\nu$ , etc. For the tube of the same length open at one end, the fundamental frequency will be  $\nu/2$ , while the frequencies of overtones will be  $3\nu/2$ ,  $5\nu/2$ ,  $7\nu/2$ ,  $9\nu/2$ , etc. Thus, the frequency of the fourth overtone of the closed tube is about 0.9 the frequency of the fourth overtone of the open tube.

I.26. If there are several antinodes in the body vibrating at an eigenfrequency, the vibration dies away irrespective of the antinode at which it is stopped (see Fig. 100b).

I.27. The frequency will be lowered.

I.28.  $\nu = 5000$  Hz and  $\lambda = 68$  mm.

I.29. The platinum string must be shorter than the steel string by a factor of 1.7.

I.30. The formula for the fundamental frequency of the string can be written as  $\nu = 2\sqrt{F/lm}$ , where  $F$  is the tensile force of the string,  $l$  is its length and  $m$  is its mass.

Thus, by increasing  $m$  we can reduce the frequency without increasing the length of the string or reducing its tension.

I.31. The pier and the dam must be longer than the wavelength of sea waves. This condition is always satisfied in practice.

I.32. In the former case, the restoring force increases due to the Coulomb attraction between the balls, i.e. the situation will be as if  $g$  has increased. Consequently, the period will be reduced. In the latter case, the Coulomb force of attraction is directed along the thread, and the period will not change (the tensile force will be slightly reduced).

I.33.  $C = 50$  pF.

I.34. The vibrator length  $l = 9.42$  m.

I.35. The capacitance should vary between 17.2 to 28.5 pF.

## Part II

## Geometrical Optics

- II.1.** The emissive power in the former case must be 11 times as large as in the latter case.
- II.2.** This technique is inapplicable since different regions of an extended source are at different distances from the point at which the intensity is measured.
- II.3.** No lens can give a strictly parallel beam in actual practice. By a “parallel” beam we always mean the rays converging or diverging at a small angle.
- II.4.** 1000 lm.
- II.5.** 2500 lx.
- II.6.**  $\Phi = 1260$  lm,  $E = 11$  lx.
- II.7.** In order to obtain a strictly parallel bundle of rays emerging from the hyperboloid, one should place a luminous point at its focus. This is impossible for the following reasons: (a) any radiator has a definite (although small) size, and (b) any optical system has errors in forming images; for this reason, the focus of a real system is not a geometrical point. The fact that the wave nature of light leads to the deviation of light beam from parallelism due to diffraction (see Chap. 8) is even more significant.
- II.8.** 28 600 cd/m<sup>2</sup>.
- II.9.**  $5 \times 10^8$  cd/m<sup>2</sup>.
- II.10.** The illuminance is 50 lx at the middle of the table and 25.6 lx at the edge of the table.
- II.11.** The absorbed luminous flux  $\Phi_a = 800$  lm and the transmitted flux  $\Phi_r = 700$  lm. The reflection coefficient  $\rho = 0.25$ , the transmission coefficient  $\tau = 0.35$ .
- II.12.**  $\Phi_e = 650$  lm. Since the transmitted flux in this case is  $\Phi_r = 0$ ,  $\Phi_a = 350$  lm.
- II.13.**  $E = 20\,000$  lx,  $L = 4330$  cd/m<sup>2</sup>.
- II.14.** 12 730 cd/m<sup>2</sup>.
- II.15.** In order to simplify calculations, we can assume that the Sun is a disc diameter  $d = 1.4 \times 10^6$  km and of constant luminance  $L = 1.5 \times 10^9$  cd/m<sup>2</sup>. Then  $I = 2.25 \times 10^{27}$  cd,  $E = 10^5$  lx.
- II.16.**  $L = 2.26 \times 10^8$  cd/m<sup>2</sup>.
- II.17.**  $I = 225$  cd.
- II.18.** The image will be sharp if the aperture is small enough (but not too small, Fig. 179). As can be seen from Fig. 177 the image is inverted.
- II.19.** The angles formed by rays with the normal to the mirror after the reflection are the same for all the rays of the bundle in view of the law of reflection.
- II.20.** 45°.
- II.21.** The angle of incidence is determined from the condition  $\tan \varphi = n$ .
- II.22.** Let us analyze Figs. 182*a* and *b*. Let the angle of incidence of ray *CB* in Fig. 182*b* be equal to the angle of incidence of ray *AB* in Fig. 182*a*, i.e.  $i_1 = i$ . According to the law of reflection,  $i = i'_1$ , and hence,  $i_1 = i'$ . Applying the law of reflection once again, we obtain  $i_1 = i'_1$ , and since  $i_1 = i$ , we have  $i'_1 = i$ . But this means that the direction of ray *BA* in Fig. 182*b* coincides with the direction of ray *BA* in Fig. 182*a*, which proves the reversibility of light rays in reflection.
- II.23.** According to the principle of reversibility of light rays, such a system cannot be realized.
- II.24.**  $n = 1.07$ .
- II.25.**  $i_{\text{refr}} = 33^\circ$ .
- II.26.** For the displacement  $l$  of the ray, we derive the following formula:  $l = d \sin(\varphi - r)/\cos r$ . In the given case,  $l = 3.45$  mm.
- II.27.** (b) As a result of refraction of light rays at the interface between water and air, they get into the observer's eye. The observer “sees” the coin in the continuation of the rays passing through air. (c) The explanation is the same as in case (b). (d) In a desert, heated air lies immediately above the hot sand, and a layer of colder air with a large refractive index is above

this layer. The light ray  $n$  is bent due to the nonuniformity of the refractive index of air. For this reason, when it gets into the eye of an observer, it seems that it emerges from point  $A'$ . The observer sees simultaneously the top of tree  $A$  and its "reflection"  $A'$  which creates the illusion of a tree on the bank of a lake. (e) Due to refraction of light, the fish sees the tree on the bank strongly displaced upwards and inclined. Due to the total internal reflection, the image of the diver is obtained above the surface of water.

$$\text{II.28. } \frac{n}{a} + \frac{n'}{a'} = \frac{n' - n}{r}.$$

$$\text{II.29. For } 1/a = 0, \text{ we have (see Ex. II.28) } a' = f' = \frac{n'r}{n' - n}. \text{ Similarly, for } 1/a' = 0, a = f = \frac{nr}{n' - n}. \text{ Hence } \frac{f}{f'} = \frac{n}{n'}.$$

$$\text{II.30. } 133 \text{ cm (here } 1/r_2 = 0).$$

$$\text{II.31. } a' = 66.7 \text{ cm, } \beta = 0.667, \text{ and } \gamma = 1.5.$$

$$\text{II.32. The image is virtual, } a' = 40 \text{ cm, } \beta = 2 \text{ and } \gamma = 0.5.$$

$$\text{II.33. The image is inverted, } a' = 60 \text{ cm, } \beta = 2.$$

$$\text{II.34. Hint. Make use of the basic formula for a thin lens (see (10.3.6)).}$$

$$\text{II.35. Hint. Use the constructions presented in Figs. 217-221, 210 and 214.}$$

$$\text{II.36. } f = 50 \text{ cm.}$$

II.37. Hint. Construct the images of several points lying on the segment and connect the obtained points by a solid line.

$$\text{II.38. } 2\beta.$$

II.39. Hint. Make use of formulas (11.4.2) and (11.4.4). (a) The image is virtual and erect,  $\beta = 3$ ,  $\gamma = 1/3$ . (b) The image is real and inverted,  $\beta = 1.5$ ,  $\gamma = 2/3$ . (c) The image is real and inverted,  $\beta = 0.6$ ,  $\gamma = 5/3$ .

II.40. Using formula (11.4.1), we obtain

$$f' = \frac{x'}{\beta} = \frac{45}{3} = 15 \text{ cm.}$$

Laying off this distance from the foci, we find the position of principal planes  $HH'$  and  $H'H'$  of the system (Fig. 426). The rear principal plane lies within the system, while the front plane is in front of the system.

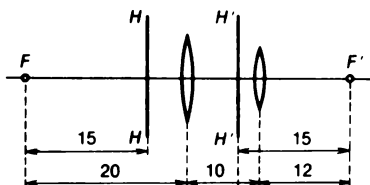


Fig. 426.

To Exercise II.40.

II.41. A camera with a telescopic objective makes it possible to obtain photographs with a large scale for a small length of the camera.

II.42. The aperture ratio is proportional to the square of the optic power of a lens.

II.43. According to formula (11.11.5), we have the following expression for the illuminance of the image:

$$E' = \frac{\Phi'}{\sigma'} = \frac{L'A}{a'^2} \approx \frac{LA}{a'^2}.$$



Substituting into this formula the expression for the luminance of the object (formula (8.10.1)), we obtain the illuminance of the image:

$$E' = \frac{\pi L d^2}{4 f^2} = \frac{q E}{4} \left( \frac{d}{f} \right)^2 = \frac{0.70 \times 40}{4} \left( \frac{1}{2.5} \right)^2 = 1.12 \text{ lx.}$$

**II.44.** Using the formula for the illuminance  $E'$  from the previous problem, we obtain

$$E' = \frac{0.95 \times 8 \times 10^8 \times \pi \times 2^2}{4 \times 5000^2} = 95 \text{ lx.}$$

**II.45.** Figure 427 shows that

whence 
$$h = (x + f) \tan \alpha = (x' + f') \tan \alpha',$$

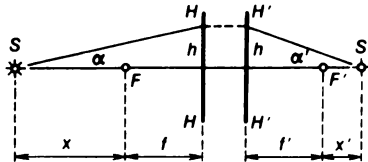
$$\gamma = \frac{\tan \alpha'}{\tan \alpha} = \frac{x + f}{x' + f'}.$$

But according to formulas (11.2.1), (11.4.1) and (11.4.2), we have

$$f' = f, \quad x = \frac{f}{\beta}, \quad x' = f' \beta = f \beta.$$

Thus,

$$\gamma = \frac{f/\beta + f}{f\beta + f} = \frac{1}{\beta}.$$



**Fig. 427.**  
To Exercise II.45.

**II.46.** *Hint.* Make use of formulas (11.2.1), (11.4.1) and (11.4.2).

**II.47.** 20.6 cm.

**II.48.**  $f = 60$  cm.

**II.49.** The losses constitute 87% without coating, and 33% with a coating.

**II.50.**  $\times 2$ .

**II.51.** 1.6 mm.

**II.52.** In the given case,  $a' = -D$ , where  $D$  is the distance of normal vision, the minus sign indicating that the object and its image are on the same side of the lens. From the lens equation, we obtain

$$\frac{1}{a} = \frac{1}{f} + \frac{1}{D}.$$

Substituting the value of  $1/a$  into the formula for the magnification of a magnifier, we obtain

$$N = \frac{\varphi'}{\varphi} = \frac{D}{a} = \frac{D}{f} + 1.$$

**II.53.** In the simplest case, it is sufficient to slightly pull out the eyepiece.

**II.54.**  $\times 500$ .

**II.55.** As a result of reflection at the first prism, the left and right change places. As a result of reflection at the second prism, up and down change places, while the objective inverts the image completely. Thus, the system as a whole gives an erect image. (The presence of the eyepiece does not change anything since it produces an erect image.)

**II.56.**  $\times 50$ .

**II.57.**  $\times 64$ .

**II.58.** A lens with larger focal length must be taken as an objective. The magnification of the telescope is  $\times 5$ . The lenses must be arranged at 18 cm from each other.

**II.59.**  $f_1 = 24$  cm,  $f_2 = 2$  cm.

**II.60.** The size of the screen should be  $1.5 \times 2.25$  m<sup>2</sup>. It must be placed at 6.5 m from the objective. The condenser must be placed immediately in front of a slide, i.e. at 26 cm from the objective, and its diameter must be about 11 cm. The distance from the condenser to the source will be 11.7 cm, the focal length of the condenser, 80 mm.

**II.61.** 0.01 s.

## Part III

## Physical Optics

**III.1.** See Sec. 14.3.

**III.2.** See Sec. 13.5.

**III.3.** See Sec. 13.5.

**III.4.** (a) In transmitted light,  $r_{10} = 7.55$  mm, in reflected light,  $r_{10} = 7.75$  mm. (b)  $\lambda = 546$  nm, (c)  $l = 1.75$   $\mu$ m, (d)  $N = 7$ .

**III.5.** About 14 m.

**III.6.**  $\lambda_2 = \lambda_1(1000/999) = 589.6$  nm for  $N = 1000$ .

**III.7.** The equidistant lines are parallel to the edge of the wedge. The separation between neighbouring maxima or minima is 2.7 mm. As the angle between the plates increases, the width of the lines decreases.

**III.9.** 80 lines, the number of lines is proportional to the thickness  $d$  and does not depend on the size of the plates.

**III.10.** The separation between adjacent maxima is  $h = D\lambda/l$ .

**III.11.**  $S_1S_2 = 0.18$  mm.

**III.12.** The width of an interference fringe is about 32 mm. As the angle of the biprism decreases, the fringes become narrower. With increasing distance from the screen, the fringes become wider.

**III.13.** The fringes become narrower.

**III.14.**  $r_1 = \sqrt{D\lambda}$ ,  $r_2 = \sqrt{2D\lambda}$  ( $\lambda$  is negligibly small as compared to  $D$ ).

**III.15.** 3.14 mm<sup>2</sup>.

**III.16.**  $\lambda = 540$  nm.

**III.17.** For  $\lambda < d$  in the first order and for  $\lambda < d/n$  in the  $n$ -th order.

**III.18.** The maximum integer less than or equal to  $d/\lambda$ . For the numerical example, 19 orders.

**III.19.** At least 10 lines.

**III.20.**  $\lambda_m/\lambda_n = n/m$ . (a) The ultraviolet lines of 300 and 200 nm, (b) the infrared line of 900 nm. If the observation is carried out with the help of a photographic plate, such an overlapping may spoil the spectrogram. For the visual observation, this overlapping is immaterial.

**III.21.** 2.9, 5.7 and 8.6°.

III.22. (a) 0.1 and 0.2 mm, (b) 100 mm, (c) in the fourth order, (d) the dispersion does not depend on wavelength, 6 nm/mm, 3 nm/mm.

III.23. The explanation should be sought in diffraction by the slit formed by the eyelids of the screwed up eye and by the grating formed by the eyelashes.

III.24. 6 mm (since the average wavelength of light can be taken as 500 nm for visual observations).

III.26.  $5.6 \times 10^8$  km.

III.27. 0.3 mm.

III.28. See Sec. 12.4.

III.29.  $0.3 \mu\text{m}$ . *Remark.* A further decrease in the focal length of the objective is associated with a decrease in its diameter, i.e. with a decrease in its angular resolution. Therefore, the value  $0.3 \mu\text{m}$  obtained in the solution determines the minimum size distinguishable through the microscope (the maximum resolving power of the microscope).

III.30. About 2 cm. In actual practice, letters must be much larger because the blackboard is not black enough and the letters written in chalk are not very distinct.

III.31. This distance is about 120 km for the eye, and about  $1/4$  km for the telescope.

III.32. It will appear white (see Table 10) in the former case and green in the latter case. The first method is known as *addition of colours* (additive colour), and the latter, *subtraction of colours* (subtractive colour).

III.33. Bright-yellow, dark-yellow, bright-yellow, yellow and black.

III.34. See Secs. 19.8, 19.9 and 19.13.

III.35.  $v_{\text{max}} = 7.7 \times 10^5$  m/s.

III.36.  $\lambda = 527$  for sodium,  $\lambda = 275$  nm for tungsten and  $\lambda = 233$  nm for platinum.

III.37. (1) Positive, (2) about 0.8 nm, (3) the result will not change since the work function for different metals are small in comparison with  $h\nu$  for  $X$ -radiation.

III.38. About 35.

III.39. See Sec. 19.12.

III.40. (a) Dark-dark, (b) dark-yellow-green, (c) dark-dark, and (d) dark-dark.

## Part IV

## Atomic and Nuclear Physics

IV.1.  $8e$ .

IV.2. 9800 km/s.

IV.3. 4.1 eV.

IV.4. (a) 20 cm, (b) 10 cm.

IV.5. 7.1 cm.

IV.6. 41 cm.

IV.7. The period of revolution does not depend on the particle velocity, see Sec. 23.7.

IV.8.  $4.35 \times 10^{-8}$  s.

IV.9.  $m/m_0 \approx 1.002$ , 3 and 2000 for the electron and 1.000001, 1.001 and 2 for the hydrogen atom.

IV.10.  $5.5 \times 10^{-12}$  kg.

IV.12. 1.005, 7.1.

IV.13.  $3.5 \times 10^{-11}$  g.

IV.14. 6.4 mm.

IV.15. 1.5 mg.

IV.16. 12, 10.1, and 1.8 eV.

IV.17. Absorption lines with the wavelengths of 122, 103, 97 nm, etc. (the Lyman series) emerge as a result of illumination by the light emitted by gaseous hydrogen at a low temperature (atoms are in the ground state).

IV.18.  $F_e = 9 \times 10^{-8}$  N,  $F_g = 4 \times 10^{-47}$  N.

IV.19. Transforming the first equation from (22.16.5) and squaring both its sides, we obtain

$$p_e^2 c^2 + m_e^2 c^4 = (h\nu)^2 + (h\nu')^2 + m_e^2 c^4 - 2h\nu \cdot h\nu' + 2h(\nu - \nu')m_e c^2$$

or

$$p_e^2 = \left( \frac{h\nu}{c} \right)^2 + \left( \frac{h\nu'}{c} \right)^2 - 2 \left( \frac{h\nu}{c} \right) \left( \frac{h\nu'}{c} \right) + 2hm_e(\nu - \nu').$$

Comparing this expression for  $p_e^2$  with the second equation from (22.16.5), we obtain

$$- \frac{2h^2}{c^2} \nu\nu' + 2hm_e(\nu - \nu') = -2 \frac{h^2}{c^2} \nu\nu' \cos \vartheta,$$

$$\frac{h}{c^2} \nu\nu'(1 - \cos \vartheta) = m_e(\nu - \nu'),$$

$$\frac{h}{m_e c} (1 - \cos \vartheta) = c \frac{(\nu - \nu')}{\nu\nu'} = \frac{c}{\nu'} - \frac{c}{\nu} = \lambda' - \lambda = \Delta\lambda.$$

Thus

$$\Delta\lambda = \frac{2h}{m_e c} \sin^2 \frac{\vartheta}{2} = 2\lambda_0 \sin^2 \frac{\vartheta}{2}.$$

IV.20. See Sec. 22.17.

IV.21.  $6 \times 10^8$ .

IV.22. See Sec. 22.6.

IV.23.  $2.4 \times 10^{-12}$  A.

IV.25. In 1 min.

IV.26.  $E = 5 \times 10^5$  V/m and  $B = 0.23$  T for the  $\alpha$ -particle and  $E = 10^6$  V/m and  $B = 0.0053$  T for the  $\beta$ -particle.

IV.27.  $3.7 \times 10^{10}$  electrons per second.

IV.28.  $18.5 \times 10^{10}$   $\alpha$ -particles per second.

IV.29.  $18 \text{ mm}^3$ .

IV.30.  $2.3^\circ \text{ C}$ .

IV.31.  $0.38 \text{ mg}$ .

IV.32. (a)  $4.5 \times 10^9$  years, (b)  $1.2 \times 10^9$  years.

IV.33.  $\tau = 4.4 \times 10^{-8}$  s (the frequency  $\nu = 2.3 \times 10^7$  Hz).

IV.34. About 110 MeV, 550 revolutions.

IV.35. (1)  ${}^1_1\text{H} + {}^1_1\text{H} \rightarrow {}^2_1\text{H} + p$ , (2)  ${}^2_1\text{H} + {}^2_1\text{H} \rightarrow {}^3_2\text{He} + n$ , (3)  ${}^6_3\text{Li} + p \rightarrow 2{}^4_2\text{He}$ , (4)  ${}^{27}_{13}\text{Al} + {}^4_2\text{He} \rightarrow {}^{30}_{15}\text{P} + p$ .

IV.36. The repulsive forces acting between an  $\alpha$ -particle and a nucleus are proportional to  $Z$ .

IV.37.  $6.4 \times 10^{-7}$  mg.

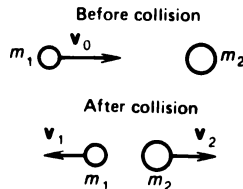


Fig. 428.  
To Exercise IV.40

IV.38. (1)  ${}^2_1\text{H} + \gamma \rightarrow p + n$ , (2)  $p + n \rightarrow {}^2_1\text{H} + \gamma$ , (3)  ${}^9_4\text{Be} + \gamma \rightarrow 2{}^4_2\text{He} + n$ ,  
 (4)  ${}^{14}_7\text{N} + n \rightarrow {}^{14}_6\text{C} + p$ , (5)  ${}^9_4\text{Be} + {}^2_1\text{H} \rightarrow {}^{10}_5\text{B} + n$ .

IV.39. 9.6%.

IV.40. Let the velocity of a ball of mass  $m_1$  before collision be  $v_0$ , and the ball of mass  $m_2$  be at rest. After the collision, the velocities of the balls are  $v_1$  and  $v_2$  respectively (Fig. 428). The total momentum of the balls before collision is  $m_1 v_0$  (the momentum of the second ball was zero). The total momentum after collision is  $m_1 v_1 + m_2 v_2$ . According to the momentum conservation law, the following equality must be observed:

$$m_1 v_0 = m_1 v_1 + m_2 v_2.$$

Going over from the vector to scalar equation, we obtain

$$m_1 v_0 = m_1 v_2 - m_1 v_1$$

(vectors  $m_1 v_1$  and  $m_2 v_2$  have opposite directions, and hence the magnitude of their sum is equal to the difference of the magnitudes of the vectors). On the other hand, according to the energy conservation law, we have

$$m_1 v_0^2 = m_1 v_1^2 + m_2 v_2^2.$$

Solving these equations together for  $v_1$  and  $v_2$ , we obtain the fraction of the initial energy supplied by the ball  $m_1$  to  $m_2$ :

$$\frac{m_2 v_2^2}{m_1 v_0^2} = \frac{4m_1 m_2}{(m_1 + m_2)^2}.$$

If we put  $m_1 = 1$  (the neutron mass in amu) and  $m_2 = A$  (the mass of the nucleus in amu), we obtain on the right-hand side  $4A/(1 + A)^2$ .

IV.41. After a collision between a neutron and a proton, the mean energy of the neutron is equal to half the initial energy:  $E_{m(n=1)} = E_0/2$ . After  $n$  collisions,

$$E_{m(n)} = E_0 \left( \frac{1}{2} \right)^n.$$

IV.42. 20 collisions.

IV.43. The radioactive atoms formed over many half-lives will have decayed by the end of exposure.

IV.44. Moving across the plate, the particle loses a fraction of its energy to ionize and excite the atoms of the medium. As a result, its velocity decreases and the trajectory is bent by the magnetic field stronger. Consequently, the particle moved downwards and carries a positive charge.

IV.45. See Sec. 24.8.

IV.46. 1.7 MeV, 7.3 MeV.

IV.47. 234.1165 amu.

IV.48. The change in energy by 10 eV corresponds to the change in mass by  $10^{-8}$  amu. If the accuracy of measurement is  $10^{-6}$  amu the change in mass by  $10^{-8}$  amu cannot be noticed.

IV.49. Electrons.

IV.50. 900 kW.

IV.51. About 8% of the liberated energy is evolved in the moderator.

IV.52. The reflector returns to the reactor a fraction of escaping neutrons, which leads to a decrease in the critical mass.

IV.53. (a) Increases, (b) is constant, (c) decreases.

IV.54. Since the power of the reactor is constant, the multiplication factor is equal to unity, i.e. one from 2.5 fission neutrons causes a new fission. 1.25 neutrons are captured without causing fission. The remaining 0.25 neutrons (i.e. 10% of all fission neutrons) escape from the reactor.

**IV.56.** During the life of one generation of neutrons, the number of fission acts increases  $k$ -fold, and during the life of  $n$  generations, by a factor of  $k^n$ . The number of generations required for an  $a$ -fold increase in the number of fission is  $n = \log a / \log k$ ,  $n = 5500$ , and  $t = 550$  s.

**IV.58.** Let a proton of mass  $m_p$ , kinetic energy  $W_k$  and momentum  $p$  impinge on a stationary target proton (the hydrogen atom nucleus). Let us consider in general form the reaction of formation of one or several particles (we shall denote them all by  $X$ ) with a total mass  $m_X$ :  $p + p \rightarrow p + p + X$ . The minimum initial energy required for the reaction (known as the energy threshold of the reaction) is spent when after a collision all particles in the final state move as a single whole with the same velocity  $v$  in the direction of the initial momentum and when they carry together this momentum (by the momentum conservation law). Then no additional energy is required for their flying apart. Applying the energy and momentum conservation laws (22.6.4) and (22.7.3) of relativistic mechanics, we obtain

$$(W_k)_{\text{threshold}} + m_p c^2 + m_p c^2 = (2m_p + m_X) \frac{c^2}{\sqrt{1 - \beta^2}}, \quad (1)$$

$$p = (2m_p + m_X) \frac{\beta c}{\sqrt{1 - \beta^2}}, \quad (2)$$

where  $\beta = v/c$ , and the total energy of the bombarding proton  $W_p = (W_k)_{\text{threshold}} + m_p c^2$  is connected with its momentum  $p$  through the well-known relation (see (22.7.4))

$$W_p^2 = p^2 c^2 + m_p c^4. \quad (3)$$

From relation (1), we obtain

$$\begin{aligned} W_p^2 &= [(W_k)_{\text{threshold}} + m_p c^2]^2 = \left[ \frac{(2m_p + m_X) c^2}{\sqrt{1 - \beta^2}} - m_p c^2 \right]^2 \\ &= \frac{(2m_p + m_X)^2 c^4}{1 - \beta^2} + m_p^2 c^4 - 2m_p c^4 (2m_p + m_X) \frac{1}{\sqrt{1 - \beta^2}}. \end{aligned}$$

From (2) and (3) we have

$$W_p^2 = p^2 c^2 + m_p^2 c^4 = \frac{(2m_p + m_X)^2 \beta^2 c^4}{1 - \beta^2} + m_p^2 c^4.$$

Equating these expressions, we get

$$\frac{(2m_p + m_X)^2 c^4 (1 - \beta^2)}{1 - \beta^2} = 2m_p c^4 (2m_p + m_X) \frac{1}{\sqrt{1 - \beta^2}},$$

or

$$\frac{1}{\sqrt{1 - \beta^2}} = \frac{2m_p + m_X}{2m_p}.$$

Then it follows from (1) that

$$\begin{aligned} (W_k)_{\text{threshold}} &= \frac{(2m_p + m_X)^2 c^2}{2m_p} - 2m_p c^2 \\ &= 2m_X c^2 \left( 1 + \frac{m_X}{4m_p} \right). \end{aligned}$$

Case (a):  $p + p \rightarrow p + p + \pi^0$ ,  $m_X = m_{\pi^0} = 135 \text{ MeV}/c^2$ . This gives

$$(W_k)_{\text{threshold for } \pi^0} = 2m_{\pi^0}c^2 \left( 1 + \frac{m_{\pi^0}}{4m_p} \right) = 2.072 \times 135 \text{ MeV} \approx 280 \text{ MeV}.$$

Case (b):  $p + p \rightarrow p + p + \bar{p} + p$ ,  $m_X = 2m_p = 1877 \text{ MeV}$ . This gives

$$(W_k)_{\text{threshold for } \bar{p}p} = 4m_p c^2 \left( 1 + \frac{1}{2} \right) = 6m_p c^2 \approx 5630 \text{ MeV}.$$

**IV.59.** 68 MeV.

**IV.60.** 0.8 MeV.

# Conclusion

The study of physical phenomena not only acquaints us with a wide range of facts, but also brings to light the laws obeyed by these phenomena, and thus makes it possible to exercise a control over physical phenomena. Moreover, by finding the laws establishing a quantitative relationship between various aspects of phenomena we can also find the causal relationship between them. This results in the emergence of the physical theory which provides an insight into the observed regularities and at the same time predicts hitherto undiscovered phenomena. By creating the conditions stipulated by the theory, we can experimentally verify the correctness of these predictions. If the predictions of the theory are confirmed experimentally, this justifies our faith in the correctness of theoretical concepts. Otherwise, we have to reconsider the theory and supplement or modify it, or even look for a new explanation for the experimentally observed processes and phenomena. This is the way of continuous development of science, based on experiment and subjected to experimental verification. We owe this supremacy over nature of science alone.

The growth of each branch of physics entails important technical applications. A knowledge of the laws of mechanics of solids, liquids and gases is behind all the achievements of modern civil engineering, from majestic skyscrapers to supersonic jet planes. Each component in these structures is manufactured keeping in mind the physical laws. The laws of thermal phenomena form the basis of all achievements in thermal engineering, from the primitive steam engines to the modern internal combustion engines and jet engines of enormous power. All the existing electrical equipment is the result of technical implementation of the basic laws of electrodynamics and the electromagnetic induction effect. Modern radio engineering with its wide range of applications in radio communication, broadcasting, television, radio astronomy, etc., right from the storm indicator and the first radiotelegraph invented by Popov, is closely connected with the development of electrodynamics which predicted and confirmed the existence of electromagnetic waves. Finally, the relatively new field of nuclear power engineering is completely based on the first experimental investigations in atomic physics and theoretical concepts lying at the root of all advances in atomic physics.



However, we are indebted to physics not only for these invaluable technical applications. To a considerable extent, our understanding of the real world and our outlook at large are influenced to a considerable extent by the progress in physics. Conversely, real progress in physics is possible only through a correct materialistic outlook. The evolution in science was always accompanied by a tough battle of philosophical views, which ultimately is a struggle of materialism with the idealistic approach towards nature.

Materialism considers physical phenomena to exist objectively, independently of the perceiving subject and to be governed by objective laws. According to the concepts of idealists, the outer world is linked inseparably with the perceiver or governed by laws that cannot be determined in their final form. Idealism is basically opposed to the scientific outlook which aims mainly at finding the laws of nature and developing concepts which produce a picture of the real world and which envisage a control over physical phenomena. Hence blatant idealism which refutes the objective nature of the world and frequently proclaims it as unknowable did not and could not meet with the approval of scientists. Modern idealism, however, assumes quite refined forms, exploiting the difficulties frequently encountered by natural scientists.

The evolution of physics has encountered several examples where idealistic philosophy led some physicists to physically untenable conclusions like the rejection of the concept of atoms and molecules, the prediction of the so-called "thermal death" of the Universe, an erroneous interpretation of the theory of relativity as a confirmation of the tentative and subjective nature of science, etc.

The development of atomic and nuclear physics has led to the discovery of peculiar laws governing the behaviour of elementary particles constituting nuclei, atoms and molecules. These laws, known as the laws of quantum mechanics, are quite different from the laws established from the observation of the motion of much larger objects which are dealt with in classical mechanics and astronomy. It should be recalled that individual atoms, their constituent nuclei and electrons, as well as other particles on atomic and subatomic scale are called microparticles in physics. The laws governing the behaviour of such particles are called the laws of the microcosm. Bodies formed by an enormous number of microparticles constitute the so-called macrocosm which includes not only objects on "human scale", but also giant bodies like stars, planets and other astronomical objects. Using this terminology, we can formulate a more precise statement, the laws of ordinary mechanics of the macrocosm are too coarse to describe the behaviour of microparticles. On the contrary, the laws of quantum mechanics are applicable to microparticles as well as to ordinary classical phenomena. In the latter

case, we obtain results identical to those given by classical mechanics and confirmed by experiment.

Thus, classical mechanics must be considered as the first approximation to the laws of the real world, sufficient for studying the motion of macroscopic objects. Quantum mechanics is a more refined and closer approximation, i.e. a more general theory that includes classical mechanics each time when we consider macroscopic objects whose mass is much larger than that of microparticles. It should be observed, however, that even in physics of the macroworld, some phenomena can be explained only with the help of quantum mechanics. These include the phenomena like superconductivity of solids and superfluidity of liquid helium.

Every new achievement in modern physics confirms the correctness of the materialistic approach and provides a deeper understanding of physical phenomena. The large-scale application of the atomic (to be more precise, nuclear) energy and the production of large quantities of new (transuranic) elements that do not exist in nature, as well as many other achievements of modern nuclear physics serve as a reliable practical verification of intricate aspects of the modern physical theory. Further progress can be made only by following the materialistic approach which has justified the entire process of scientific evolution. This progress is so rapid and has acquired such dimensions that it is rightfully called a revolution in science and technology.

# Index

- Aberration, 236ff
  - chromatic, 240f
  - spherical, 236ff
- absorption of light, resonance, 415
- acceleration, free-fall, 27
- accelerator(s), 459ff, 467, 523ff
  - colliding-beam, 524ff
  - cyclotron, 460
  - fixed-target, 524f
- accommodation, 251
- acoustics, 47ff
  - physical, 48
- aerial, 125, 129
- albedo, 176
- amplification, 145
- analysis, luminescence, 373
- angle,
  - critical, of total internal reflection, 198
  - of deflection, 202
  - of incidence, 190f, 196
  - of reflection, 97
  - of refraction, 190, 196
  - of view, 252
- annihilation, 475
- antibaryon(s), 528
- antideuteron(s), 510
- antimatter, 510
- antineutrino, 510
  - electron, 510f
  - muon, 511
- antineutron(s), 509
- antinode(s), 110, 115
- antinucleus(i), 510
- antiparticle(s), 506ff
- antiproton(s), 508
- aperture, 218
- aperture ratio, 249
- astigmatism, 212, 239f
- atom(s), 388ff
  - energy levels of, 410, 413
  - hydrogen, 419ff
  - many-electron, 423f
  - mass of, 403f
  - structure of, 388ff
- automation, 154
- baryon(s), 513, 528
- baryon charge, 509, 513, 529
- beam,
  - light, 186
  - paraxial, 210
- beauty, 530
- biprism, 273f
- blackbody, 358f, 361
- brightness, 360
  - of illuminated surface, 176f
  - of image, 243f, 265f
  - of source, 169ff
- Brownian movement, 389
- bubble chamber, 514
- camera,
  - pinhole, 187
  - obscura, 187
- candela, 168
- capacitance, 61, 66
- charge-to-mass ratio, 397
- charm, 530
- Chladni figures, 115ff
- choroid, 251

- chromatography, 334
- cinema projector, 247
- circuit,
  - LC-, 65, 74f, 87
  - LR-, 65
  - RC-, 65
- clock,
  - pendulum, 33
  - water, 33
- clock paradox, 518
- coefficient,
  - absorption, 174, 342
  - reflection, 174, 342
  - transmission, 174, 342
- coherence, 273, 418
- collimator, 349
- colour(s), 195, 334ff, 341f
  - complementary, 338f
  - of light, 275
  - saturation of, 345
  - simple, 336
  - of thin films, 276f
- condenser, 246, 248
- condition, quantization, 437
- consonance, 56
- constant,
  - Avogadro's, 389f
  - Planck's, 367, 402, 429, 435
  - Rydberg, 353
  - solar, 96
- cornea, 251
- cosmic rays, 519ff
- counter, Geiger-Müller, 449
- current,
  - photoelectric, 365, 368
  - saturation of, 365
- Damper(s), 32
- de Broglie wavelength, 434ff
- decay,
  - alpha, 455
  - beta, 455, 499f
- detector(s), 146
- diaphragm, 241
- diascope, 246
- diffraction, 98ff
  - of light, 282ff, 287, 315
  - Fresnel's interpretation of, 290f
- diffraction grating, 294ff
  - preparation of, 297
  - as a spectral instrument, 296
- diffraction pattern, 289
- dioppter, 230
- dispersion of light, 195, 334ff
- dissonance, 56
- echo sounder, 84
- effect(s),
  - binaural phase, 121f
  - Compton, 431
  - of light, 157ff, 363ff
  - photochemical, 374
  - photoelectric, 157f, 363ff
- efficiency, economic, 360
- eigenfrequency(ies), 118
- electric charge, elementary, 391ff
- electron(s), 388, 498ff, 506
  - mass of, 398f
  - velocity dependence of, 398
  - recoil, 431f
- electron-positron pair(s), 475
- electron shell(s), 423
- emission,
  - of electromagnetic waves, perfect, 124f
  - of light,
    - induced, 415
    - spontaneous, 415
  - photoelectric, 366
  - of sound, perfect, 119f
  - thermionic, 366
- energy,
  - internal, 401
  - kinetic, 24, 28, 31, 111, 128, 401, 429, 480
  - nuclear, 480ff
  - potential, 24, 28, 31, 111
    - of electron, 437
    - of photon, 432
  - radiant, 162
  - rest, 401
  - total, 31, 401f
- epidiascope, 248
- episcope, 248
- equation,
  - of spherical mirror, 219
  - of thin lens, 210, 212

- experiment(s),
  - Becquerel's, 441
  - Compton's, 430f
  - M. Curie's, 442
  - Hertz', 129ff, 132, 134, 141
  - Lebedev's, 129, 131
  - Millikan's, 392f
  - Newton's, 335f
- eye, 250
- eye front chamber, 251
- eye lens, 251f
- eyepiece, 256, 259
- feedback, 73
- field,
  - gravitational, 503
  - of nuclear forces, 503
- field of vision, 241
- flavour, 530
- fluorescence, 322, 341, 372
- flux,
  - luminous, 162, 172f, 175
  - radiation, 162
- focal length, 206, 233, 249
- focus(i),
  - front, 209, 228, 232
  - principal, 206, 228
  - rear, 209, 228, 232
- force(s),
  - anharmonic periodic, 39
  - centrifugal, 23
  - of gravity, 23
  - harmonic, 41, 118
  - Lorentz, 395
  - nuclear, 479, 502
  - restoring, 23f, 28
  - tensile, 23
- formant(s), 55
- formula,
  - Balmer, 353f
  - de Broglie, 434f, 437
  - Einstein's, 367
  - Thomson, 64f
- frequency,
  - fundamental, 41f, 52, 113f
  - vibrational, 28, 113
- friction, 30
- gamma-quantum(a), 475ff, 517, 520
- gamma-rays, 503
- generator(s), quantum, 415ff
- hadron(s), 528f, 513, 533
  - quark structure of, 537
- half-life, 456
- harmonic(s), 41ff
  - amplitude of, 44
  - phase, 44
- hologram, 301, 311
- holography, 299
  - application to optical interferometry, 310ff
- holographic reconstruction of image, 305f
- holographic recording, 302, 306
- hydroacoustic detection, 83
- hyperon(s), 516, 537
- hyperopia, 252
- iconoscope, 148
- illuminance, 165ff, 176, 242
- image, 204
  - diminished, 227, 229
  - erect, 229
  - inverted, 227
  - magnified, 229
  - real, 212ff
  - virtual, 212ff, 216, 228f
- inductance, 61, 66
- interaction(s),
  - electromagnetic, 511, 534
  - electroweak, 544
  - gravitational, 511
  - nuclear, 511
  - strong, 511, 534
  - weak, 511, 534
- interference of light, 159f, 272ff, 315
- interference pattern, 104ff, 278
- intermediate boson(s), 542f
- isotope(s), 403ff
  - separation of, 405
- isotopic family, 529

- lamp,
  - gas, 341
  - incandescent, 360
- laser(s), 273, 324, 416ff
  - ruby, 417
- law(s),
  - Boyle's, 390
  - conservation, 528
    - of baryon charge, 529
    - of energy, 402, 431, 477, 480
    - of mass, 402
    - of momentum, 431, 524
  - Einstein's, 400ff, 480
    - of annihilation and pair formation, 476f
  - of geometrical optics, 182ff
  - of illumination, 166ff
  - Kirchhoff's, 357ff
  - Lenz's, 62
  - Newton's second, 399
  - of photoelectric effect, 364ff
  - of radioactive displacement, 455f
  - of reflection of light, 188ff
  - of refraction of light, 188ff
  - spectral, 353
- lens(es), 205, 222f
  - aperture of, 242
    - relative, 242
  - converging, 213ff, 227
  - diverging, 213ff
- lepton(s), 513, 542ff
- light,
  - absorption of, 346
  - induced emission of, 415
  - linearly polarized, 317
  - natural, 316
  - reflection of, 284, 346f
  - refraction of, 284f
  - spectral composition of, 340f
  - wavelength of, 189
- light beam, coherent, 300
- lighting engineering, 171
- light quantum(a), 367
- light year, 333
- locator, sound, 122
- loudness, 48, 95
- lumen, 168
- luminous flux, 243, 368
- luminous intensity, 165ff
- lux, 168
- luxometer, 179
- magnification, 254
  - angular, 224, 235
  - linear, 223ff, 234
  - transverse, 223
- magnifier, 254ff
  - magnification of, 255
- masking, 344
- mass, critical, 486
- mass number, 404
- mass radiator, 134
- mass spectrograph, 395f
- Mendeleev's Periodic System of Elements, 424ff
- meson(s), 502f, 513, 528, 538
  - pi-, 503ff
  - mu-, 506
- method,
  - double-exposure, 314
  - electroacoustic, 48
  - holographic, opposing light beam, 308
  - of labelled atoms, 494
- microscope, 256ff
  - magnification of, 257
  - optical length of, 257
  - resolving power of, 258
- mirror,
  - convex, 219
  - plane, 216
  - spherical, 217ff
- model,
  - of atom,
    - nuclear, 407ff
    - planetary, 421
  - quark, 538ff
    - of hadron structure, 533
- modulation, 144
  - telegraph, 144
  - telephone, 144
- moment of inertia, 30
- motion,
  - aperiodic, 32
  - periodic, 13
- myopia, 252

- neutrino(s), 499ff
- neutron(s), 388, 467, 479, 498ff
  - fission, 484, 487
  - properties of, 468ff
- Newton's rings, 277ff
- node(s), 110
- noise, 55ff
- nuclear power, 494f
- nucleon(s), 502
- nucleus(i), 465ff
  - structure of, 477ff
- objective(s), 247ff, 250, 257
- optical axis, 206, 315
  - auxiliary, 206, 227
  - principal, 206, 217, 227
- optical centre, 206, 252
- optical glass, 221
- optical instrument(s), projection-type, 246
- optical power, 230
- optical system(s), 232ff
- optic nerve, 251
- optics, physical, 272
- oscillation(s),
  - amplitude of, 15ff
  - antiphase, 13ff
  - of charge, 128
  - of current, 128
  - damped, 31
  - electric, 58ff
    - undamped, 70f
  - forced, 14
  - free, 13f, 31
    - undamped, 31
  - frequency of, 18, 20
  - harmonic, 18ff, 42, 57
  - natural, 31
  - of pendulum, 23f
    - damping of, 30ff
    - free, 30
  - period of, 15ff, 20
  - periodic, 42
  - phase of, 22
  - plane, 25
  - synphase, 22
  - undamped, 31
- oscillator(s), 125ff
  - continuous wave (CW), 129, 135
  - electric, 125
    - frequency of, 127
    - simple, 27, 33
      - period of, 33
    - valve, 73f, 134
  - oscillatory circuit, 61ff
  - oscillatory system(s), 13ff
    - self-excited, 70ff
  - oscillogram(s), 15ff, 45, 50
  - oscillograph,
    - cathode-ray, 59f
    - light-beam, 16f
- parsec, 333
- particle(s),
  - alpha-, 407ff, 444, 452ff, 465f, 473, 478
  - beta-, 444, 452ff
  - elementary, 498
- path difference, 107
- pendulum, 14ff
  - oscillations of, free, 30
  - physical (compound), 23
  - simple, 23ff, 27
    - amplitude of, 25
    - length of, 26f
    - period of, 25ff
  - torsional, 30
    - period of, 30
- period, 13
- phase shift, 21f
- phosphorescence, 322, 372
- photocell, 158, 369f
  - barrier-layer, 370
- photoelectric multiplier, 448
- photographic camera, 248
- photography, 299, 375ff
- photoluminescence, 363, 371
- photometer, 177f
- photometry, 162ff
- photon(s), 368, 498ff
  - properties of, 427ff
- phototelegraphy, 147f
- pitch, 48f
- polarization of light, 315ff
- polarization plane, 317
- polaroid, 318

- positron(s), 473ff, 399, 506
- principle, Huygens', 283ff
  - Fresnel's interpretation of, 286
- prism, 336
  - erecting, 199
  - reflecting, 199
- proton(s), 388, 465, 479, 498ff
- pupil, 251
- pyrometry, 361f
  
- quantum mechanics, 433ff
- quantum number(s), 530
  - principal, 438
  
- radiation,
  - alpha-, 449ff, 473
  - beta-, 443ff, 472
  - gamma-, 443ff, 451f, 473
  - infrared, 321f
  - intensity of, 196
  - isotropic, 165
  - radioactive, 441
    - properties of, 451ff
  - ultraviolet, 321f
  - uniform, 165
- radio, 80
  - communication, 142
  - invention of, 141f
- radioactive decay, 455ff, 480
- radioactive family of uranium, 458
- radioactivity, 441ff
  - application of, 459
  - artificial, 472
- radio astronomy, 152
- radio compass, 141
- radio echo, 151
- radiogeodesy, 149
- radiolocation, 83f
- radionavigation, 149
- radiophysics, 153
- ray, light, 182, 185
  - reversibility of, 192f
- reaction(s),
  - fission, 483f
    - application of, 488
    - chain, 484
    - nuclear, 465ff, 481f
    - chain, 468
      - neutron-induced, 470ff
      - thermonuclear, 483
  - reactor(s), nuclear, 490ff
  - receiver, crystal, 147
  - reflection, total internal, 197f
  - reflector(s), 173
  - refraction,
    - in lens, 205ff
    - in plane-parallel plate, 200f
    - in prism, 202f
  - refractive index, 194ff, 336ff
    - absolute, 194
    - relative, 195
  - resistance, 61f
  - resolving power, 291ff, 304
  - resonance, 34ff
    - acoustic, 51f
    - broad, 36
    - effect of friction on, 36
    - electric, 67ff
    - with many frequencies, 117f
    - sharp, 36
  - rest mass, 399, 401f
  - retina, 251, 257, 265
  - reverberation, 98
  - rigidity, 29f
  
  - scale of electromagnetic waves, 132, 327f
  - scatterer(s), 173
  - scintillation counter, 448
  - sclera, 250
  - solar batteries, 371
  - sound, 46
  - sound ranging, 83
  - sound recording, 53
  - sound reproduction, 53
  - source(s),
    - coherent, 107
    - of light, 204
    - point, 163ff, 183
  - spectral analysis,
    - qualitative, 356
    - quantitative, 356
  - spectral line(s),
    - absorption, 357
    - emission, 357
    - Fraunhofer, 357f



spectral series,  
     Balmer, 353, 414  
     Lyman, 354  
     Paschen, 354, 414  
 spectrograph, 350  
 spectroscopy, 350  
 spectroscopy, 349  
 spectrum(a), 335, 349ff, 398  
     absorption, 357f  
         of atoms, 357  
         of liquids, 357  
         of solids, 357  
     acoustic, 51  
     band, 350  
     continuous, 350, 352  
     electromagnetic, 320f  
     emission, 354f, 357  
     harmonic, 45, 50  
     line, 350, 352  
     optical, analysis of, 412ff  
 speed of light, 312ff  
     measurement of, 330ff  
 spinthariscopes, 448  
 steradian, 154  
 stereoscope, 267ff  
 Stokes' shift, 371  
     physical meaning of, 373  
 strangeness, 530  
 sweep, 60  
  
 telecontrol, 149  
 telegraph, wireless, 80  
 telephone, 59  
 telescope, 258ff  
     applications of, 261  
     Galilean, 260  
     Keplerian, 259  
     Lomonosov, 267  
     magnification of, 260f  
     reflector, 262ff  
 tembre, 49f  
 theorem, Fourier, 42, 52  
 theory,  
     Einstein's, of relativity, 502  
     electromagnetic, of light, 132, 318f  
     Maxwell's, 129  
     of oscillations, 75  
     photochemical, of vision, 379

tone, 48, 52  
 transformation(s),  
     photochemical, 364  
     radioactive, 455  
 transuranic elements, 493  
  
 umbra, 183  
 unit(s),  
     of charge, 394f  
     of energy, 394f  
     of mass, 394  
  
 velocity, 24, 28  
     of light, 82, 132  
     of sound, 82  
 vibration(s),  
     acoustic, 46ff  
     anharmonic, 49  
     elastic, 27ff  
         period of, 29  
     of elastic bodies, 111  
     forced, 33ff  
         amplitude of, 35  
         period of, 34  
     free, of a string, 112  
     infrasonic, 46  
     torsional, 29f  
     ultrasonic, 46  
 vision, stereoscopic, 268  
 vitreous body, 251  
  
 wave(s), 13ff  
     acoustic, 100  
         interference of, 108f  
     blast, 79  
     circular, 97  
     coherent, 104, 109  
     compression, 90  
     elastic, 90  
     electromagnetic, 123ff, 138, 161, 179, 319  
         scale of, 132ff  
     emission of, directional, 104  
     energy transfer by, 93ff  
     frequency of, 104  
     intensity of, 101, 162

interference of, 103ff  
light, 161  
    monochromatic, 300  
longitudinal, 88ff  
mechanical, 79ff  
radio,  
    diffraction of, 150  
    propagation of, 149  
reference, 302f  
reflected, 135f  
reflection of, 96ff  
refracted, 135ff  
running, 110  
seismic, 79  
shear, 90  
shock, 314  
sinusoidal, 86  
spherical, 94f, 97, 282

standing, 105ff, 135, 437  
surface, 91ff  
transverse, 85ff, 318  
    velocity of propagation of, 81f  
wavelength, 86, 101f, 104, 133, 337  
wave front, 182, 282  
Wilson cloud chamber, 443, 445ff, 474,  
    484, 514  
work function, 366

X-rays, 389f, 428ff, 445, 503  
    diffraction of, 434f  
    effects of, 324  
    nature of, 326f  
    origin of, 326f  
X-ray tube, 325f, 393

## **To the Reader**

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:  
Mir Publishers  
2 Pervy Rizhsky Pereulok  
I-110, GSP, Moscow, 129820  
USSR







## ABOUT THE PUBLISHERS

Mir Publishers of Moscow publishes Soviet scientific and technical literature in twenty five languages including all those most widely used. Titles include textbooks for higher technical and vocational schools, literature on the natural sciences and medicine, popular science and science fiction. The contributors to Mir Publishers' list are leading Soviet scientists and engineers from all fields of science and technology. Skilled translators provide a high standard of translation from the original Russian. Many of the titles already issued by Mir Publishers have been adopted as textbooks and manuals at educational establishments in India, France, Cuba, Syria, Brazil, and many other countries.

Mir Publishers' books in foreign languages can be purchased or ordered through booksellers in your country dealing with V/O "Mezhdunarodnaya Kniga".

# **ELEMENTARY TEXTBOOK ON PHYSICS**

**Edited by Landsberg**

## **CONTENTS**

### **Volume 1**

#### **Mechanics. Heat. Molecular Physics**

Kinematics. Dynamics. Statics. Work and Energy. Curvilinear Motion. Motion in Noninertial Reference Systems and inertial Forces. Hydrostatics.

Aerostatics. Fluid Dynamics. Thermal Expansion of Solids and Liquids. Work. Heat. Law of Energy Conservation. Molecular Theory. Properties of Gases. Properties of Liquids. Properties of Solids. Transition from Solid to Liquid State. Elasticity and Strength. Properties of Vapours. Physics of the Atmosphere. Heat Engines.

### **Volume 2**

#### **Electricity and Magnetism**

Electric Charges. Electric Field. Direct Current. Thermal Effect of Current. Electric Current in Electrolytes. Chemical and Thermal Generators. Electric Current in Metals. Electric Current in Gases. Electric Current in Semiconductors. Basic Magnetic Phenomena. Magnetic Field. Magnetic Field of Current. Magnetic Field of the Earth. Forces Acting on Current-Carrying Conductors in Magnetic Field. Electromagnetic Induction. Magnetic Properties of Bodies. Alternating Current. Electric Machines: Generators, Motors and Electromagnets.

### **Volume 3**

#### **Oscillations and Waves. Optics. Atomic and Nuclear Physics**

Basic Concepts. Mechanical Vibrations. Acoustic Vibrations. Electric Oscillations. Wave Phenomena. Interference of Waves. Electromagnetic Waves. Light Phenomena: General. Photometry and Lighting Engineering. Basic Laws of Geometrical Optics. Application of Reflection and Refraction of Light to Image Formation. Optical Systems and Errors. Optical Instruments. Interference of Light. Diffraction of Light. Physical Principles of Optical Holography. Polarization of Light. Transverse Nature of Lightwaves. Electromagnetic Spectrum. Velocity of Light. Dispersion of Light and Colours of Bodies. Spectra and Spectral Regularities. Effects of Light. Atomic Structure. Radioactivity. Atomic Nuclei and Nuclear Power. Elementary Particles. New Achievements in Elementary Particle Physics.

**ISBN 5-03-000225-2**

**ISBN 5-03-000223-5**